UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS

Identification of novel biomarkers and candidate genes associated to lipid traits: improving the lipid metabolism knowledge base

"Documento Definitivo"

Doutoramento em Biologia
Biologia de Sistemas

Marta Sofia da Silva Correia

Tese orientada por:
Margarida Gama Carvalho
Mafalda Bourbon

Documento especialmente elaborado para a obtenção do grau de doutor

2022

UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS

Identification of novel biomarkers and candidate genes associated to lipid traits: improving the lipid metabolism knowledge base

Doutoramento em Biologia

Biologia de Sistemas

Marta Sofia da Silva Correia

Tese orientada por:

Margarida Gama Carvalho

Mafalda Bourbon

Júri:

Presidente:

- Rui Manuel dos Santos Malhó, Professor Catedrático e Presidente do Departamento de Biologia Vegetal da Faculdade de Ciências da Universidade de Lisboa

Vogais:

- Doutora Maria Paula Borges de Lemos Macedo, Professora Associada com Agregação, NOVA Medical School da Universidade Nova de Lisboa
- Doutor Matthias Erwin Futschik, Professor Auxiliar Convidado, Faculdade de Medicina e Ciências Biomédicas da Universidade do Algarve
- Doutora Maria Luísa Mourato de Oliveira Marques Serralheiro, Professora Associada com Agregação, Faculdade de Ciências da Universidade de Lisboa
- Doutora Margarida Henriques da Gama Carvalho, Professora Auxiliar, Faculdade de Ciências da Universidade de Lisboa (Orientadora)

Documento especialmente elaborado para a obtenção do grau de doutor

Projeto financiado pela FCT (PD/BD/114387/2016)

2022

Marta Sofia da Silva Correia foi bolseira de doutoramento (PD/BD/114387/2016) no âmbito do programa doutoral BioSys PD65-2012 (FCT/PD/00065/2012) da Faculdade de Ciências da Universidade de Lisboa, com financiamento pela Fundação para a Ciência e Tecnologia (FCT) do Ministério da Ciência, Tecnologia e Ensino Superior. Este trabalho foi também financiado pela FCT através dos apoios ao centro de investigação do BioISI UIDB/04046/2020 e UIDP/04046/2020.

Agradecimentos

Queria agradecer à minha orientadora, a Professora Margarida Gama Carvalho, por me acolher no seu grupo de investigação e me ter dado a oportunidade de enveredar pelos caminhos da bioinformática. Obrigada por todo o seu apoio, motivação e tempo investido, sem os quais não teria sido possível para mim terminar o doutoramento. Mesmo à distância conseguiu arranjar sempre um espacinho no horário para falar comigo, esclarecer as minhas dúvidas e permitir que eu não perdesse o fio condutor. Obrigada por ter apostado em mim!

À minha coorientadora, Doutora Mafalda Bourbon, por me conceder a oportunidade de trabalhar com o Estudo Português de Hipercolesterolemia Familiar, pelo acesso aos dados que possibilitaram este trabalho e pelo esclarecimento das dúvidas mais clínicas. Ao grupo de investigação cardiovascular do INSA pelo fornecimento de todos os dados necessários e pela resposta às questões que me foram surgindo sobre o *dataset* e a FH em geral.

Ao Professor Francisco Pinto, por todo o apoio dado e disponibilidade para as minhas dúvidas de bioinformática e estatística, sempre com simpatia.

Aos meus colegas do grupo RNASysBioLab da FCUL, por toda a ajuda prestada, partilha de ideias e dicas de código, assim como pela boa disposição, almoços animados e momentos de lazer. Obrigada a todos, aos que passaram pelo gabinete de computacional (Daniel Olivença, Tânia, Lúcia, Marcelo, André, Marina, Gonçalo, Sofia, Fábio, Daniel Eleutério, Guillem), e aos colegas do laboratório, principalmente à Cláudia e ao Mark.

Aos colegas de doutoramento, em particular à terceira edição do BioSys, pelos momentos passados, tanto académicos como de lazer, ao longo deste percurso. Ao Daniel, que é meu colega de curso desde a licenciatura (2009!), Ana Rita, Diana, João, Márcia, Mariana, Madalena, Margarida, Rafael e Teresa.

Às amizades que me trouxe esta minha aventura por Lisboa, em particular à Catarina, por todos os momentos de convívio, longas conversas existenciais, e por seres uma excelente guia turística.

A todos aqueles que de alguma forma contribuíram para uma estadia mais fácil numa cidade até então desconhecida e me ajudaram a adaptar consoante as circunstâncias.

Por último, queria deixar um agradecimento muito especial à minha família, principalmente aos meus pais e ao meu irmão, por todo o apoio prestado ao longo destes últimos anos e pela confiança que sempre depositaram em mim. Sem o suporte deles nunca teria sido possível concluir esta etapa. Obrigada!

Abstract

FH, the most common monogenic dyslipidaemia, is characterised by increased circulating LDL-C levels leading to premature cardiovascular disease when undiagnosed or untreated. Current guidelines support genetic testing in patients fulfilling clinical diagnostic criteria and cascade screening of their family members. However, about half of clinical FH patients do not present pathogenic variants in the known disease genes (*LDLR, APOB, PCSK9*), and these most likely suffer from polygenic hypercholesterolaemia, which translates into a relatively low yield of genetic screening programs. This project aimed to identify new biomarkers able to improve the distinction between monogenic and polygenic profiles.

Using a machine-learning approach in a paediatric dataset, tested for disease causative genes and investigated with an extended lipid profile, we developed new models that classify FH patients with higher specificity than currently used methods. The best performing models incorporated parameters absent from the common FH clinical criteria, which rely only on TC and LDL-C. A hierarchical clustering analysis of the same dataset showed that the study population can be clearly divided in three groups of dyslipidaemic individuals, showing the complexity of the dyslipidaemic biological context and the need of an integrative and multidisciplinary approach for biomarker selection. Both clustering and modelling analysis have revealed that the extended lipid profile contains important biomarkers.

The exploration of lipid metabolic pathways associated with the identified biomarkers allowed us to identify a set of related genes. Using additional information from public databases, including gene expression data, associated GWAS and GO terms, we defined a universe of lipid-related genes and molecular interactions relevant for the dyslipidaemic context and future genetic studies. All this information was used to establish a new lipid knowledge base available online.

The obtained results can be applied to improve the yield of genetic screening programs and decrease the associated costs, and also provide novel contributions to our understanding of dyslipidaemias.

Keywords: familial hypercholesterolaemia, biomarkers, extended lipid profile, machine-learning based methods, lipid knowledge base.

Resumo

A hipercolesterolemia familiar (FH) é a dislipidemia monogénica mais comum, com uma prevalência de 1/250 em heterozigotia e 1/300 000 em homozigotia, e resulta da presença de variantes patogénicas nos genes *LDLR, APOB e PCSK9*. Apesar de ser relativamente comum em relação a outras doenças de causa monogénica, permanece subdiagnosticada e a maioria das pessoas afetadas não se encontram medicadas. Esta doença caracteriza-se principalmente pelos níveis elevados de colesterol total e LDL-C no sangue, promovendo a acumulação de colesterol no interior dos vasos sanguíneos, formação de placas de ateroma e desenvolvimento de aterosclerose. Sendo uma doença silenciosa, os sintomas só costumam surgir na ocorrência de um evento cardiovascular agudo (por exemplo, enfarte do miocárdio ou acidente vascular cerebral). Atualmente, é recomendado o estudo genético em indivíduos com critérios clínicos para FH, bem como o rastreio em cascata para os familiares. No entanto, metade dos indivíduos com hipercolesterolemia grave não apresenta uma variante patogénica nos genes associados à FH; nestes casos suspeita-se que estes indivíduos sofram de uma forma poligénica de hipercolesterolemia com possível modulação por fatores ambientais, como por exemplo um estilo de vida pouco saudável. Deste modo, os programas de rastreio genético revelam pouca eficiência, ao mesmo tempo que os critérios clínicos responsáveis pela seleção de indivíduos para estudo genético são altamente sensíveis mas pouco específicos.

Este projeto teve como principal objetivo a identificação de novos biomarcadores capazes de distinguir os doentes com dislipidemia monogénica dos doentes com dislipidemia poligénica, contribuindo para uma melhor seleção de indivíduos para estudo genético, bem como para uma visão mais integrada do metabolismo lipídico e das relações entre os vários intervenientes (nomeadamente, genes, proteínas e metabólitos). Esta integração do conhecimento lipídico permitiu a identificação de um grupo restrito de genes candidatos, os quais poderão ser utilizados em futuros estudos moleculares direcionados à dislipidemia. Para além disso, esta análise integrativa contribuiu para a construção de uma nova base de conhecimentos lipídicos, específica para o metabolismo dos lípidos e dislipidemia, disponível online para a comunidade científica.

Com a aplicação de uma abordagem metodológica baseada em *machine learning* num *dataset* pediátrico, constituído por um perfil lipídico alargado e com os indivíduos estudados para os genes associados à FH, foram desenvolvidos novos modelos capazes de classificar os indivíduos com FH com uma especificidade consideravelmente superior em comparação com os métodos tradicionais. Os modelos com melhor performance, considerando um *ranking* com

o “top 10” dos modelos, incorporam parâmetros lipídicos ausentes dos critérios clínicos para FH atualmente utilizados, os quais apenas consideram os valores de colesterol total e LDL-C. Adicionalmente, modelos utilizando os mesmos parâmetros lipídicos que os critérios clínicos tradicionais (colesterol total e LDL-C) revelaram uma maior especificidade, mantendo uma boa sensibilidade, na classificação de uma amostra de 50 indivíduos do PFHS-ped, face ao uso de *cut-offs* rígidos desses mesmos parâmetros. Estes resultados demonstram o elevado potencial de técnicas de *machine learning* para desenvolvimento de modelos de classificação que permitam melhorar a seleção de indivíduos para estudo genético, e consequentemente, contribuir para um diagnóstico precoce.

Uma análise de *clustering* hierárquico aplicada sobre o mesmo *dataset* revelou que a população em estudo pode ser dividida em três grupos distintos de doentes dislipidémicos, em vez das duas classes comumente utilizadas com base nos resultados dos estudos genéticos (FH+, presença de critérios clínicos e variante patogénica num dos três genes associados à doença; ou FH-, quando há presença de critérios clínicos mas não há identificação de uma variante associada). O terceiro grupo de indivíduos corresponde a um perfil misto que compreende algumas características de um perfil FH+ e outras mais próximas de um perfil FH-. Adicionalmente, esta análise permitiu identificar diferentes padrões lipídicos entre os indivíduos, nomeadamente a associação entre um perfil FH+ e níveis mais elevados de parâmetros lipídicos envolvidos no metabolismo das LDL/apoB, enquanto um perfil FH- foi associado com níveis mais elevados de parâmetros envolvidos no metabolismo dos triglicéridos. Uma maior contribuição poligénica foi também encontrada em indivíduos FH-, tendo em conta um score de risco poligénico associado aos níveis de LDL-C e baseado num painel de seis SNPs. Estes resultados mostram a complexidade do contexto biológico da dislipidemia, e a necessidade de uma abordagem integrativa e multidisciplinar para a seleção de novos biomarcadores.

À semelhança dos modelos desenvolvidos para classificação dos indivíduos, a análise de *clustering* revelou que o perfil lipídico alargado contém biomarcadores importantes. Desta forma, a disponibilidade de um leque mais diversificado de parâmetros lipídicos pode ser uma mais-valia na procura de novos biomarcadores. Resumidamente, os parâmetros relacionados com o metabolismo dos triglicéridos e das LDL/apoB mostraram contribuir para uma melhor distinção entre indivíduos, de acordo com a análise de *modelling* e *clustering*. Os modelos com melhor performance, considerando o “top 10”, são maioritariamente compostos por um grupo de parâmetros que inclui os triglicéridos, LDL-C, apoB, apoA-I, apoC-III e LDL1. Os primeiros quatro destes parâmetros estão também presentes nos primeiros cinco lugares do *ranking* dos

parâmetros com contribuição estatisticamente significativa para a distribuição dos indivíduos em três grupos, na análise de *clustering* hierárquico, onde também se encontra o VLDL. Ambas as análises baseadas em métodos de *machine learning* mostraram que o metabolismo das lipoproteínas é uma das principais vias metabólicas envolvidas na dislipidemia.

Posteriormente, a exploração de vias metabólicas lipídicas associadas aos biomarcadores identificados durante as análises anteriores permitiu identificar um conjunto de genes de interesse na área da dislipidemia. O primeiro passo foi a realização de uma pesquisa direcionada na plataforma *Wikipathways*, onde os biomarcadores identificados posteriormente foram utilizados como palavras-chave. Adicionalmente, outros termos foram considerados como palavras-chave, incluindo "metabolismo lipídico", "hipercolesterolemia" e "aterosclerose", com base na revisão da literatura. Esta pesquisa resultou na seleção de 14 vias metabólicas de interesse, e todos os genes envolvidos nestas vias foram utilizados para estabelecer uma lista de 466 genes alvo.

Utilizando informação adicional recolhida de bases de dados públicas, incluindo dados de expressão génica, traços de fenótipo/doença (a partir de resultados de estudos de associação ampla do genoma) e termos de ontologia genética, foi definida uma nova base de conhecimento lipídico constituída pelo universo de 466 genes alvo relacionados com o metabolismo dos lípidos e dislipidemia, e interações moleculares relevantes para o contexto biológico da dislipidemia e futuros estudos moleculares. Considerando o perfil fenotípico e funcional dos genes alvo, foi estabelecido um grupo mais restrito de 41 genes denominado de "genes candidatos", os quais são definidos pela associação com pelo menos um dos nove traços fenotípicos previamente selecionados como de interesse (doseamento de lipoproteínas, colesterol total, LDL-C, HDL-C, e VLDL-C, para além de outros traços fenotípicos como hipertrigliceridemia, aterosclerose, doença cardiovascular, e doença arterial coronária) e associação com termos de ontologia genética relacionados com lípidos. O grupo de genes candidatos inclui os três genes relacionados à FH (*LDLR*, *APOB*, *PCSK9*), entre outros genes maioritariamente associados ao metabolismo das lipoproteínas e dos triglicéridos, e representa um painel de interesse para futuros estudos moleculares. Esta nova base de conhecimento lipídico foi disponibilizada online através do desenvolvimento de uma *app* (MylipidgenesKB), que permite o acesso e interação com os vários níveis de informação associados aos genes de interesse e o *download* de ficheiros, apresentando-se como um recurso útil para a comunidade científica.

Os resultados obtidos neste trabalho podem ser aplicados para melhorar os programas de rastreio genético e diminuir os custos correspondentes, bem como para um maior conhecimento do contexto biológico das dislipidemias.

Palavras-chave: hipercolesterolemia familiar, biomarcadores, perfil lipídico estendido, métodos baseados em *machine learning*, base de dados de conhecimento lipídico.

De acordo com o disposto no artigo 24º do Regulamento de Estudos de Pós-Graduação da Universidade de Lisboa, Despacho nº 7024/2017, publicado no diário da República – 2ª série – nº 155 – 11 de Agosto de 2017, foram incluídos nesta dissertação os seguintes artigos:

- **Correia, M.**, Kagenaar, E., van Schalkwijk, D.B. *et al.* Machine learning modelling of blood lipid biomarkers in familial hypercholesterolaemia versus polygenic/environmental dyslipidaemia. *Sci Rep* 11**,** 3801 (2021). https://doi.org/10.1038/s41598-021-83392-w

- **Correia, M.,** Bourbon, M., Gama-Carvalho, M. Analysis of a paediatric cohort of dyslipidaemic patients using unsupervised learning methods provides insights into the biochemical phenotypes of familial hypercholesterolemia. medRxiv, p. 2022.07.17.22277724 (2022). https://medrxiv.org/cgi/content/short/2022.07.17.22277724v1

- **Correia, M.,** Gama-Carvalho, M. A review of lipoprotein metabolism and new approaches on biomarker discovery for dyslipidaemia (2022). [in preparation]

No cumprimento do disposto da referida deliberação, a autora declara ter sido a principal executante do trabalho, tendo participado na definição das metodologias aplicadas, produção e interpretação dos resultados obtidos, assim como na redação dos manuscritos.

Contents

List of tables

List of figures

Abbreviations

7αH	cholesterol 7 α-hydroxylase
ABCA1	ATP binding cassette subfamily A member 1
ABCG1	ATP binding cassette subfamily G member 1
ABCG5	ATP binding cassette subfamily G member 5
ABCG8	ATP binding cassette subfamily G member 8
ACAT	acetyl-CoA acetyltransferase
acetyl-CoA	acetyl coenzyme A
ACMG	American College of Medical Genetics and Genomics
AIC	Akaike information criterion
ANGPTL3	angiopoietin like 3
ANGPTL4	angiopoietin like 4
apoA-I	apolipoprotein A-I
apoA-IV	apolipoprotein A-IV
apoA-V	apolipoprotein A-V
APOB	apolipoprotein B gene
apoB	apolipoprotein B
apoB-100	apolipoprotein B-100
apoB-48	apolipoprotein B-48
apoC-I	apolipoprotein C-I
apoC-II	apolipoprotein C-II
apoC-III	apolipoprotein C-III
apoE	apolipoprotein E
APOE	apolipoprotein E gene
AUC	area under the ROC curve
BMI	body mass index
BP	biological process
CD36	cluster of differentiation 36

CETP	cholesteryl ester transfer protein
CVD	cardiovascular disease
DGAT	acyl-CoA diglycerolacyltransferase
EAS	European atherosclerosis society
EOMI	Early-onset myocardial infarction
FA	fatty acids
FFA	free fatty acids
FH	familial hypercholesterolaemia
FH-	FH-negative
FH+	FH-positive
GETx	Genotype-Tisue Expression (database)
GGE	gel gradient electrophoresis
GO	gene ontology
GPIHBP1	glycosylphosphatidylinositol-anchored high density lipoprotein binding protein 1
GWAS	genome-wide association studies
HCPC	hierarchical clustering of principal components
HDL	high-density lipoprotein
HDL-C	high-density lipoprotein cholesterol
HGNC	HUGO Gene Nomenclature Committee
HL	hepatic lipase
HMG-CoA	β-hydroxy-β-methylglutaryl-CoA
HSPG	endothelial heparan sulphate proteoglycans
IBAT	ileal bile acid transporter
IDL	intermediate-density lipoprotein
KDs	knowledge (data)bases
KIV2	kringle IV type 2 – (encoding sequences)
LAL	Lysosomal acid lipase (alias symbol of LIPA)
lbLDL	large buoyant LDL

LCAT	lecithin cholesterol acyl transferase
LDL	low-density lipoprotein
LDL1	low-density lipoprotein fraction 1
LDL2	low-density lipoprotein fraction 2
LDL-C	low-density lipoprotein cholesterol
LDLR	low-density lipoprotein receptor
LDLRAP1	low-density lipoprotein receptor modular adaptor protein 1
LIPA	lipase A, lysosomal acid type
LMF1	lipase maturation factor 1
Lp(a)	lipoprotein(a)
LPL	lipoprotein lipase
LRP	low-density lipoprotein-like receptor protein
MAF	minor allele frequency
MF	molecular function
MGAT	acyl-CoA monoglycerolacyltransferase
MIDB	intermediate-density lipoprotein fraction B
MIDC	intermediate-density lipoprotein fraction C
ML	machine learning
MLPA	Multiplex Ligation-dependent Probe Amplification
MTTP	microsomal triglyceride transport protein
NGS	next generation sequencing
NPC1L1	Niemann-Pick C1-like protein 1
NPV	negative predictive values
PCA	principal components analysis
PCSK9	proprotein convertase subtilisin-like kexin type 9
PDZK1	PDZ-domain-containing protein
PFHS	Portuguese FH study
PheWAS	phenotypic-wide association study

PLTP	phospholipid transfer protein
PPV	positive predictive values
RFE	recursive feature elimination
ROC	receiver operating characteristic (curves)
SB	Simon Broome (criteria)
SCAP	SREBP cleavage activating protein
sdLDL	small dense LDL
sdLDL.Day	small dense LDL (measured by Daytona procedure)
sdLDL-C	small dense low-density lipoprotein cholesterol
SER	smooth endoplasmic reticulum
SNPs	single nucleotide polymorphisms
SRA1	scavenger receptor A-1
SRB1	scavenger receptor B-1
SRE	sterol response element
SREBP	sterol regulatory element binding protein
TC	total cholesterol
TG	triglycerides
TPM	transcript per million
VLDL	very low-density lipoprotein

Chapter 1

State of the art

1. Lipids and lipoproteins metabolism

Lipids are organic substances that comprise a heterogeneous group of compounds related to fatty acids (FA) and that include fats, oils, waxes, and other related substances. They are relatively insoluble in water and considerably soluble in organic solvents like ether, chloroform, or benzene, which make them hydrophobic [1]. Some lipids are amphipathic, being composed by a polar/hydrophilic "head" group (e.g., hydroxyl group) and a non-polar/hydrophobic hydrocarbon chain [1], [2]. Lipids play several biological functions, such as being important constituents of diet (high energy value), fat reserves (energy storage), structural elements (especially in cell membranes), players in hormone synthesis, vitamin carriers (e.g. fat-soluble vitamins as A, D or E), signalling molecules, and emulsifying agents [1]–[3].

The lipids of metabolic significance in mammals include triglycerides (TG), phospholipids and steroids (including cholesterol), together with products of their metabolism such as long-chain FA, glycerol, and ketone bodies [1], [2]. Three important tissues in the lipid metabolic network are the liver, blood, and adipose tissue. Both liver and adipose tissue are the main sites of metabolic activity, while blood functions as a transport system. In addition, other tissues, such as cardiac and skeletal muscle, are important users of FA and ketone bodies [1]. Specific liver functions in lipid metabolism comprise the following (Figure 1.1):

1. oxidation of FA to supply energy for other body functions;
2. synthesis of large quantities of cholesterol, phospholipids, and most lipoproteins;
3. synthesis of fat from proteins and carbohydrates;
4. conversion of FA into ketone bodies during fasting;
5. all processes related to cholesterol metabolism (synthesis and release into the blood, secretion of plasma cholesterol into bile, and its conversion into bile salts) [3], [4].

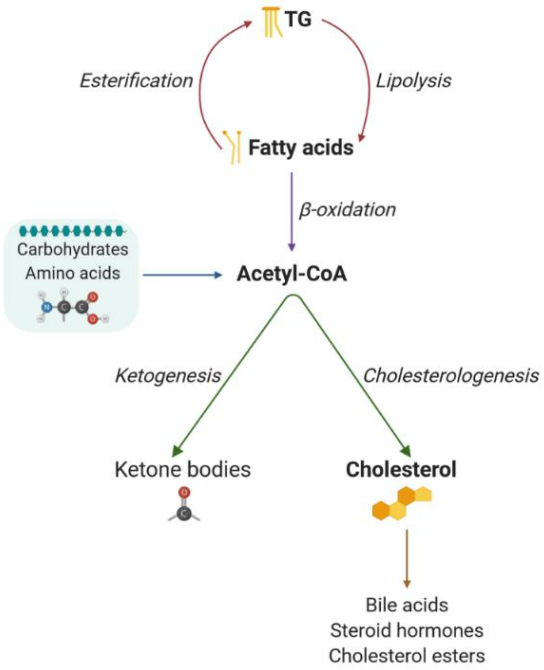


Figure 1.1. An overview of the main lipid metabolic processes and the connection to carbohydrate and amino acids metabolism. Created with BioRender.com.

1.1. Emulsification and lipolysis

TG are the most abundant lipids in diet, followed by small quantities of phospholipids, cholesterol, and cholesteryl esters. The first step in lipids digestion is the physical breakdown of fat globules into very small sizes, allowing the action of water-soluble digestive enzymes – a process called **emulsification**, which begins in the stomach through peristaltic movements. Then, most emulsification occurs in the duodenum under the activity of bile, a liver secretion that does not contain any digestive enzymes, but instead is rich in bile salts and lecithin (a phospholipid) [4]. When present in high concentration in watery fluids, bile salts form micelles – small spherical and cylindrical globules composed of 20 to 40 molecules of bile salt. The hydrophobic sterol nucleus of bile salts dissolve in the surface layer of fat globules, forming a small fat globule in the middle of the resulting micelle, with polar groups of bile salts projecting outward to cover the micelle surface. Since these polar groups are negatively charged, the entire micelle globule is able to dissolve in the digestive fluids and remain in stable solution until fat can be absorbed into the blood [4] (Figure 1.2).

Each time the diameter of fat globules decreases because of small intestine agitation, the total surface area of the fat increases as much as 1000-fold. This enables the activity of lipase enzymes that are water-soluble and can only attack fat globules on their surfaces. The most important enzyme for TG digestion is pancreatic lipase, present in high quantities in pancreatic

juice, which is responsible for splitting TG into free FA and 2-monoglycerides (Figure 1.2). Cholesteryl esters and phospholipids are hydrolysed, releasing FA, by enzyme cholesteryl ester hydrolase and phospholipase A_2, respectively. The entire process of fat digestion and splitting in smaller and simpler products is known as **lipolysis** [4].

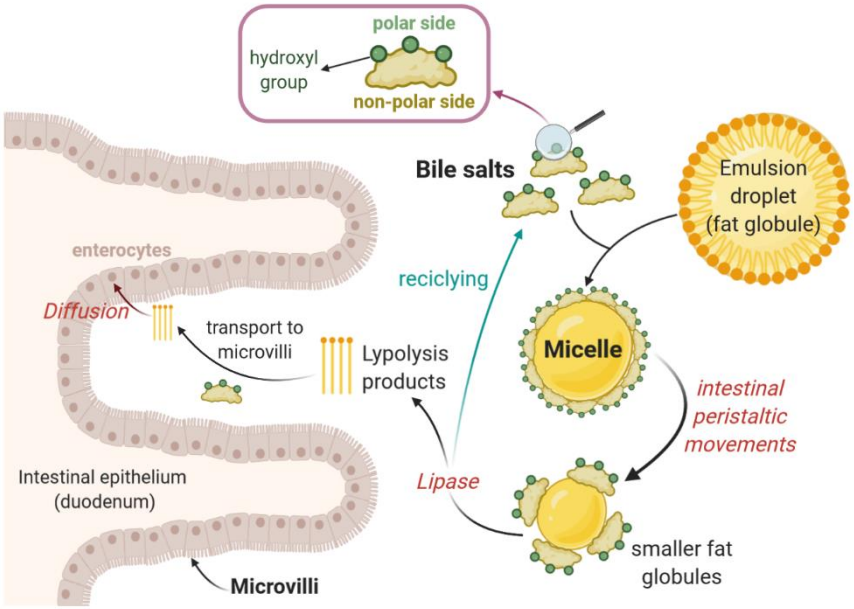


Figure 1.2. Fat digestion and absorption (by diffusion) in intestinal epithelium, with bile salts as important players in both processes. The magnification in the upper side identifies the polar and non-polar surfaces of bile salts. Created with BioRender.com.

Micelles are also involved in fat absorption, playing a role in the transport of lipolysis products to the brush borders of intestinal epithelial cells (microvilli), where those products are absorbed by diffusion in the blood. Bile salts are then released back into the chyme to be used over and over again [4].

1.2. Esterification and entry of lipids in circulation

Once inside the epithelial cell, FA and 2-monoglycerides are taken up by the smooth endoplasmic reticulum (SER), where they are mainly used to form new TG (a process called **esterification**) under the activity of enzymes acyl-CoA monoacylglycerol acyltransferase (MGAT) and acyl-CoA diacylglycerol acyltransferase (DGAT). These resynthesized TG are subsequently packaged and secreted in the form of lipoproteins (i.e., **chylomicrons**) for transport proposes [3]–[5]. The packaging process is mainly mediated by microsomal triglyceride transport protein (MTTP) and apolipoprotein B-48 (apoB-48), with apolipoprotein A-I (apoA-I) also taking part in the process. Following lipidation by MTTP, apoB-48 is lipidated by apolipoprotein A-IV (apoA-IV) [5], [6]. After being secreted, chylomicrons flow

upward through lymph ducts and enter into the circulating blood, where apoA-I and apoA-IV are exchanged for apolipoprotein C-II (apoC-II) and apolipoprotein E (apoE) from high-density lipoprotein (HDL) [3]–[5]. Lipoproteins are essential entities in fat transportation and highly important for cholesterol cell supply and excess cholesterol removal, as explained forward in this chapter.

In parallel, small quantities of short and medium-chain FA are absorbed directly into the portal vein, since they are more water-soluble in comparison to long-chain FA [3], [4].

1.3. β-oxidation

For energetic purposes, FA produced during lipolysis are split by **β-oxidation** into C_2-acetyl radicals, leading to the formation of acetyl coenzyme A (acetyl-CoA). The acetyl-CoA can enter into the citric acid cycle (also called TCA cycle) and be oxidised, which releases high amounts of energy that in turn can be used for cellular functions. β-oxidation can take place in all body cells, but it is particularly fast in hepatocytes. The liver itself cannot use all the acetyl-CoA that is formed and so it is converted in acetoacetic acid, a highly soluble acid that passes from hepatocytes into the extracellular fluid and is then transported throughout the body to be absorbed by other tissues, where it is reconverted into acetyl-CoA [4].

1.4. Cholesterologenesis

Like other biosynthetic pathways, reaction sequences related to the biosynthesis of lipids are endergonic and reductive. They use ATP as a source of metabolic energy and a reduced electron carrier, usually NADPH, as a reductant [1]. Cholesterologenesis comprises four main stages (Figure 1.3):

1. condensation of three acetate units to form a C-6 intermediate (mevalonate);
2. conversion of mevalonate to activated isoprene units;
3. polymerization of six C-5 isoprene units to form the branched-chain C-30 squalene;
4. cyclization of squalene to form the four rings of steroid nucleus, with further reactions (oxidation, removal or migration of methyl groups) to produce cholesterol [2], [7].

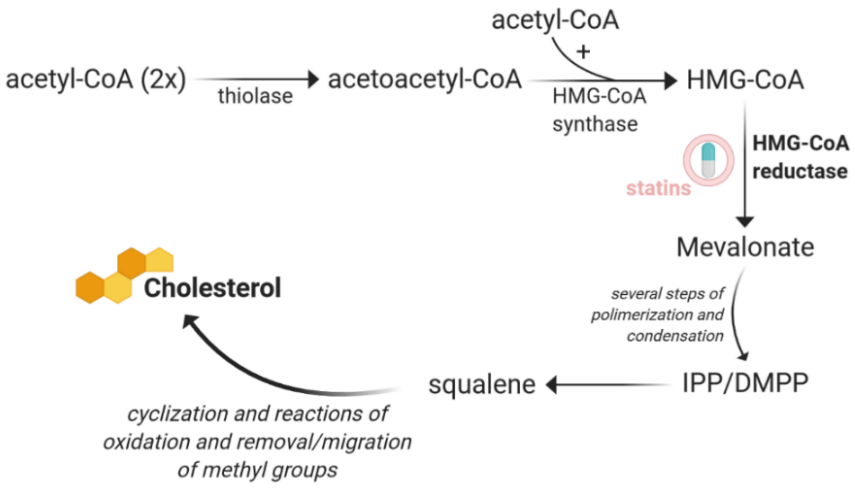


Figure 1.3. Main steps of cholesterologenesis. HMG-CoA reductase is a drug target for statins, commonly used in the treatment of hypercholesterolaemia. The inhibition of this enzyme by statins results in the inhibition of cholesterologenesis and, in turn, in lower cholesterol levels. HMG-CoA: β-hydroxy-β-methylglutaryl-CoA; IPP: isopentenyl pyrophosphate; DMPP: dimethylallyl pyrophosphate. Created with BioRender.com.

As shown in Figure 1.3, the reduction of β-hydroxy-β-methylglutaryl-CoA (HMG-CoA) to mevalonate is catalysed by the enzyme HMG-CoA reductase, an integral membrane protein that is the major point of cholesterologenesis regulation [2].

1.5. Lipoproteins: transport of lipids through the body

Cholesterol and cholesteryl esters, like TG and phospholipids, are essentially insoluble in water, yet they must be moved to the tissues in which they will be stored or consumed. Thus, body lipids are carried in the blood plasma as **lipoproteins**, macromolecular complexes of specific carrier proteins (apolipoproteins) with several combinations of phospholipids, cholesterol, cholesteryl esters and TG (Figure 1.4). The lipoprotein shell is amphipathic because its outer surface is hydrophilic (apolipoproteins), making lipoproteins water-soluble, and its inner surface (cholesterol-containing phospholipid monolayer) is hydrophobic. Adjacent to this inner surface there is a core of neutral lipids containing mostly cholesteryl esters, TG, or both. In addition, small amounts of other hydrophobic compounds (e.g., vitamin E, carotene) are carried in the lipoprotein core [2], [7].

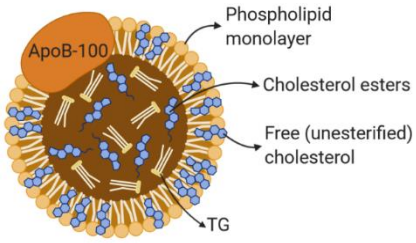


Figure 1.4. Structure of a lipoprotein (in this case, LDL), showing the phospholipid monolayer as well as the shell and the neutral core. ApoB-100 is the main apolipoprotein present in LDL. Created with BioRender.com.

1.5.1. Classes of lipoproteins

Lipoproteins fall into five major classes: HDL, low-density lipoprotein (LDL), intermediate density lipoprotein (IDL), very low-density lipoprotein (VLDL), and chylomicrons. The lower the protein/lipid ratio, the lower the density, which means that chylomicrons are the least dense and contain the highest proportion of lipids. VLDLs and chylomicrons carry mainly TG in their cores, whereas the cores of LDLs and HDLs consist mostly of cholesteryl esters. IDLs contain substantial amounts of both TG and cholesterol. Each class of lipoproteins has distinctive apolipoprotein and lipid compositions, besides of different density and size (Table 1.1) [2], [6]–[8].

Table 1.1. Major classes of plasma lipoproteins and its main properties, including density, size, main lipids composition, protein/lipid ratio and the apolipoproteins mainly present [7]–[11].

Lipoprotein	**Density (g/mL)**	**Size (nm)**	**Main lipids (% weight)**	**Protein/lipid ratio**	**Main apolipoproteins**
Chylomicron	<0.930	75-1200	TG (80-90)	1/100	B-48*, C-II, C-III, E, A-I, A-II, A-IV
VLDL	0.930-1.006	30-80	TG (40-60)	9/100	B-100*, E*, C-II, C-III
IDL	1.006-1.019	23-35	TG (20-50) and cholesterol (20-40)	19/100	B-100*, E, C-II, C-III
LDL	1.019-1.063	18-25	Cholesterol (50-60)	25/100	B-100*
HDL	1.063-1.210	5-12	Cholesterol (15-25) and phospholipids (23-30)	90/100	A-I*, A-II, C-I, C-II, C-III, E
Lp(a)	1.055-1.085	25-30	Cholesterol (20-30)	30/100	(a)*, B-100*

TG: triglycerides; VLDL: very-low density lipoprotein; IDL: intermediate density lipoprotein; LDL: low density lipoprotein; HDL: high density lipoprotein; Lp(a): lipoprotein(a)
*Most abundant apolipoprotein(s) in the particle

As shown in Table 1.1, another class of lipoproteins is considered, the lipoprotein(a) – Lp(a), which consists of a LDL particle whose apolipoprotein B-100 (apoB-100) is covalently attached

by a disulphide thioester bond to an apo(a) molecule. The apo(a) is a glycoprotein homologous to plasminogen that is highly polymorphic in size due to the number of kringle IV type 2 (KIV2)-encoding sequences [8], [11], [12]. Kringles are triple-loop structural motifs usually found in several proteases involved in blood coagulation and fibrinolysis. Apo(a) size is inversely correlated with Lp(a) density and plasma concentration, which means that individuals with smaller apo(a) isoforms have an increased apo(a) secretion [11], [13]. In fact, small apo(a) isoforms were shown to be associated with a two-fold increased risk for cardiovascular disease (CVD), in comparison to larger isoforms, and higher plasmatic Lp(a) levels are related to a higher cardiovascular risk [11], [12]. The atherogenic potential of Lp(a), which translates in the ability to contribute to the accumulation of fat within blood vessels, is justified by its interference with the fibrinolytic system, affinity to phospholipase A2, interaction with extracellular matrix glycoproteins, and binding to macrophage scavenger receptors [12].

While most lipoproteins are produced by a single type of cells (chylomicrons in enterocytes and VLDL in hepatocytes) or result from the catabolism of other lipoproteins (IDL, LDL and Lp(a) from VLDL catabolism), HDL is derived from both intestinal and hepatic tissues [6], [14], [15]. Hepatic HDL, in a nascent form, appears as a disk-shaped structure, while intestinal HDL is more spherical and varies in its protein composition. Both HDLs start to be relatively small and cholesterol poor, so that they are classified as HDL_3 (density of 1.125 to 1.210 g/mL). After interaction with lecithin cholesterol acyl transferase (LCAT) and lipoprotein lipase (LPL), cholesteryl ester content is increased and the HDL particle becomes less dense and larger, being classified as HDL_2 (density of 1.063 to 1.125 g/mL) [8], [14]. LDL particles can also be divided in different fractions according to size, density, and lipid composition [16], as explained forward in this chapter.

1.5.2. Apolipoproteins

Apolipoproteins (Table 1.2) contribute to the structural organisation of lipoproteins and determine their interactions with enzymes, extracellular lipid-transfer proteins, and cell-surface receptors [7]. The nomenclature for apolipoproteins is alphabetical. ApoA-I is the major apolipoprotein of HDL, which in turn is known as apoA-I-containing lipoprotein. Conversely, apoB-100 is present on VLDL and its remnants – LDL, IDL and Lp(a) lipoproteins. ApoB-48, a truncated protein codified by the apoB-100 gene resulting from a post-translational modification, is present in chylomicrons and its remnants. Therefore, chylomicrons, VLDL, IDL, LDL and Lp(a) are all apoB-containing lipoproteins [5], [8].

Table 1.2. Main characteristics of the major plasma apolipoproteins, including molecular weight, lipoproteins where they are present, main known metabolic functions and source [2], [8], [10], [13], [17], [18].

Apolipoproteins	Molecular weight (Da)	Lipoproteins	Metabolic function	Source
ApoA-I	28.016	HDL, chylomicrons	Removes cell cholesterol via ABC1 onto nascent HDL; cofactor LCAT; facilitates uptake of CE from HDL, LDL and VLDL by SRB1	Liver, intestine
ApoA-II	17.414		Involved in cell cholesterol efflux; suggested of hindering reverse cholesterol transport by inhibition of LCAT and CTEP	Liver
ApoA-IV	46.465		Activates LCAT; involved in chylomicron assembly and secretion	Intestine
ApoA-V	39	HDL, VLDL, chylomicrons	Stimulates proteoglycan-bound LPL	Liver (mainly)
ApoB-48	264	chylomicrons	Assembly and secretion of chylomicrons from the small intestine	Intestine
ApoB-100	540	VLDL, IDL, LDL	Assembly and secretion of VLDL from liver; binding ligand of LDL to LDLR	Liver
ApoC-I	6630	Chylomicrons, VLDL, IDL, HDL	Activates LCAT; inhibits CETP and SRB1	
ApoC-II	8900		Cofactor LPL	
ApoC-III	8800		Inhibits LPL and binding of IDL to LDLR; increases VLDL secretion	Liver (mainly), intestine
ApoE	34.145	Chylomicrons, VLDL, IDL, HDL	Ligand for uptake of chylomicron remnants and IDL by LRP and LDLR	Liver (mainly)
Apo(a)	250-800	Lp(a)	Unknown (possibly involved in wound healing and coagulation)	Liver

ABC1: ATP-binding cassette 1; CE: cholesteryl esters; LCAT: lecithin cholesterol acyl transferase; SRB1: scavenger receptor B-1; VLDL: very-low density lipoprotein; IDL: intermediate density lipoprotein; LDL: low-density lipoprotein; HDL: high density lipoprotein; LPL: lipoprotein lipase; CETP: cholesteryl ester transfer protein; LDLR: low-density lipoprotein receptor; LRP: low-density lipoprotein-like receptor protein; Lp(a): lipoprotein(a)

1.5.3. Metabolic pathways of plasma lipoproteins

Plasma lipoprotein metabolism comprises the following interrelated major pathways: exogenous (intestinal), endogenous (hepatic), intracellular LDL receptor pathway and reverse cholesterol transport [5], [8]. Chylomicrons, VLDL, and their remnants, constitute the major carriers of TG, while LDL and HDL transport most of the cholesteryl esters [5].

Only VLDL and chylomicrons are fully formed within cells by assembly in the endoplasmic reticulum, which requires the activity of MTTP. The assembled particles move through secretory pathways to the cell surface and are released by exocytosis. Thus, both VLDL and chylomicrons require specialised vesicles to traffic from endoplasmic reticulum through the

Golgi apparatus prior to being secreted [7], [15]. Conversely, other lipoproteins are generated extracellularly in the bloodstream and on cell surfaces by remodelling of secreted VLDL. There are four main modifications leading to this remodelling process as follows:

1. hydrolysis of TG and phospholipids by lipases and esterification of cholesterol by acetyl-CoA acetyltransferase (ACAT);
2. transfer of cholesteryl esters, TG and phospholipids between lipoproteins by specific lipid-transfer proteins;
3. uptake by some particles of cholesterol and phospholipids exported from cells;
4. association and dissociation of apolipoproteins from the surface of lipoproteins [7].

1.5.3.1. Exogenous pathway

Once chylomicrons enter in the systemic circulation, apoA-I and apoA-IV are exchanged for apoC-II and apoE from HDL. Then, TG within chylomicrons are hydrolysed by LPL, an enzyme that is bound to the surface of endothelial cells, and its cofactor apoC-II, producing free fatty acids (FFA) that are taken up by adipose tissue and muscle (Figure 1.5). Those FFA are then re-esterified into TG for storage purposes in adipocytes and used as energy supply by muscular cells. The chylomicrons remnants, enriched in cholesteryl esters and apoE, are quickly sequestrated by endothelial heparan sulphate proteoglycans (HSPG), within the perisinusoidal space that surrounds hepatocytes. This is followed by receptor-mediated endocytosis of remnants (Figure 1.5), through the binding of apoE to low-density lipoprotein-like receptor protein (LRP), also called chylomicron remnant receptor. HSPG also participate in this uptake step. As an alternative, apoE can be recognized and bind to a low-density lipoprotein receptor (LDLR) [5], [6], [8], [19].

After playing their role in digestion, bile acids are reabsorbed through the ileal bile acid transporter (IBAT) and recycled in the liver (Figure 1.5). Free cholesterol is synthesised in liver by HMG-CoA reductase and then it can be:

1. excreted into bile, unchanged, via ATP binding cassette subfamily G member 5 and 8 (ABCG5 and ABCG8, respectively) for excretion through stool;
2. converted to bile acids by cholesterol 7α-hydroxylase (7αH);
3. esterified by ACAT into cholesteryl esters;
4. used for lipoprotein synthesis [5], [13].

The lipase maturation factor 1 (LMF1), the glycosylphosphatidylinositol-anchored high density lipoprotein binding protein 1 (GPIHBP1), angiopoietin like 3 (ANGPTL3) and angiopoietin

like 4 (ANGPTL4) are involved in LPL mediated hydrolysis of TG, so that the right amount of FA is delivered to the right tissue at the right time [15]. LMF1 is essential to the formation of catalytically active LPL from newly synthesised polypeptides, while GPIHBP1 is involved in transporting LPL from basolateral to apical surface of endothelial cells, thus operating as platform for LPL-mediated hydrolysis of TG from chylomicron. The high affinity binding of LPL to GPIHBP1 implies that GPIHBP1 can tear LPL from HSPG. Levels of GPIHBP1 hepatic expression are low, which likely contributes to poor hepatic clearance of chylomicrons until TG hydrolysis transforms them into remnant lipoproteins. HSPG have already been implicated as receptors for lipoprotein remnants, indeed several lipid binding proteins, such as LPL, hepatic lipase (HL) and apoE, can bind to HSPG and facilitate hepatic remnant clearance [13], [19].

1.5.3.2. Endogenous pathway

Most TG in fasting state are carried by VLDL, whose synthesis is critically dependent on the amount of apoB-100, TG and cholesteryl esters in the liver. FFA are usually activated followed by oxidation and incorporation into TG or cholesteryl esters, and apoB-100 is constitutively produced in liver – its synthesis is regulated by proteolysis and not through the expression of apolipoprotein B (*APOB*) gene. As apoB-100 interacts with cholesteryl esters, it assumes a new conformation leading to decreased degradation and increased production of apoB; MTTP incorporates TG into this complex. Next, phospholipids, apoE, apoC-I, apoC-II, and apolipoprotein C-III (apoC-III) are added to mature VLDL, whose secretion requires apoB-100. Then, VLDL carry TG to adipose tissue and muscle, where TG are hydrolysed by LPL and apoC-II producing FFA, larger VLDL remnants, and after that smaller IDL remnants. IDLs are relatively enriched in cholesteryl esters and depleted in TG, they are either taken up by interaction of apoE with LDLR in liver or hydrolysed by HL, thus becoming LDL, which is the final product of VLDL catabolism (Figure 1.5). In addition, VLDLs attend as acceptors of cholesterol transferred from HDL, by mediation of cholesteryl ester transfer protein (CETP), accounting in part for the inverse relation between high-density lipoprotein cholesterol (HDL-C) and VLDL content in TG [5], [6], [8].

LDL is the major carrier of cholesterol in humans, supplying tissues with high sterol demand. Additionally, LDL is also the lipoprotein most clearly implicated in atherogenesis, as explained later in this chapter. Conversely, HDL is believed to protect tissues from the excess of cholesterol, having a major role in reverse cholesterol transport [5], [6].

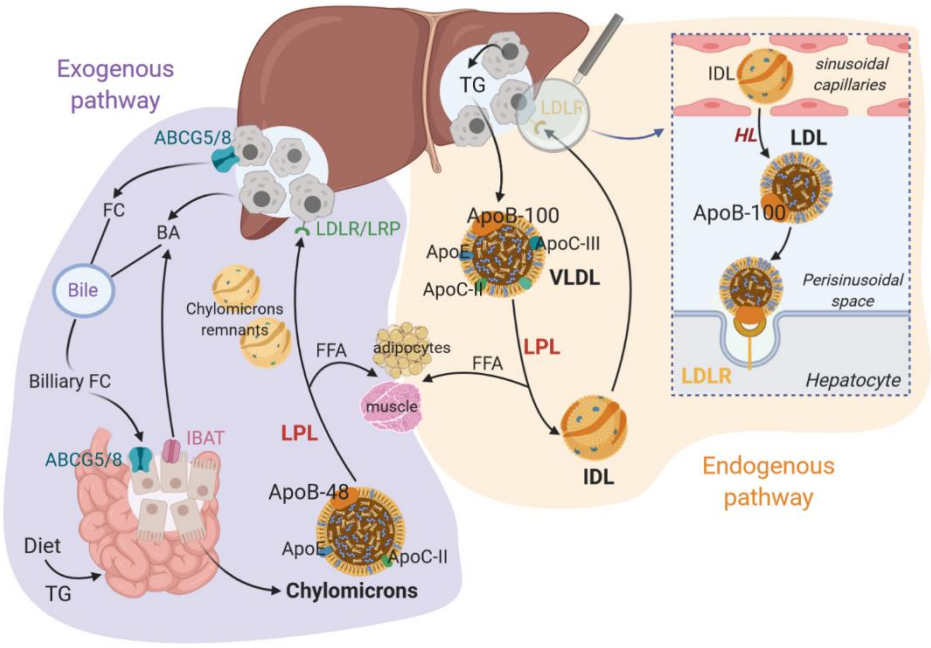


Figure 1.5. Exogenous and endogenous pathways of lipoprotein metabolism. BA: bile acids; FC: free cholesterol; FFA: free fatty acids; TG: triglycerides; LPL: lipoprotein lipase; HL: hepatic lipase. Created with BioRender.com.

LDL receptor has a higher affinity for apoE-containing lipoproteins, such as chylomicrons and VLDL remnants, than for apoB-containing lipoproteins (i.e., LDL), which might explain why LDL particles stay in circulation for more time (i.e., some days, compared to a couple of minutes or hours in other lipoproteins), thus contributing to their atherogenic potential [20]. Thus, the more VLDL remnants are removed from circulation, the less they will be converted into LDL [20], [21].

Whereas LDL receptor appears to function solely in lipoprotein metabolism, LRP has several known ligands and it may function as a multifunctional scavenger receptor, with a major function in the removal of proteinase and proteinase inhibitor complexes. It has been suggested that LRP binds not only to chylomicron remnants but also to lipases. For example, apolipoprotein A-V (apoA-V) might influence lipid homeostasis by enhancing receptor-mediated endocytosis of chylomicrons, given its association with LRP and other members of the LDL receptor family [13].

1.5.3.3. Intracellular LDL receptor pathway

After its synthesis, glycosylation, and transport to the cell surface, LDLR is directed to clathrin-coated pits where apoB of LDL particles binds with high affinity to LDLR. The receptor-ligand complexes, inside coated vesicles, are internalised by endocytosis and transported to endosomes by low-density lipoprotein receptor modular adaptor protein 1 (LDLRAP1), an important player

in LDL binding and internalisation (Figure 1.6). For degradation purposes, LDL is displaced from LDLR in the acidic environment of endosomes, allowing release of LDL into endosomes and recycling of LDLR on the cell surface. LDL is then degraded in lysosomes, where apoB undergoes proteolysis and cholesteryl esters are hydrolysed producing free cholesterol and FFA. Free cholesterol decreases the activity of HMG-CoA reductase and LDLR, by inhibition of the sterol regulatory element binding protein (SREBP) pathway. In alternative, LDLR interacts and complexes with a protease called proprotein convertase subtilisin-like kexin type 9 (PCSK9), which is secreted by the liver. This LDLR-PCSK9 complex is internalised via clathrin-mediated endocytosis and then routed to lysosomes for degradation (Figure 1.6) [5], [10].

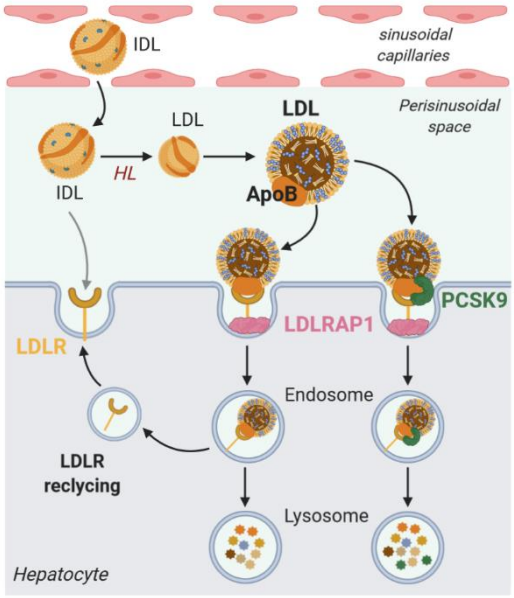


Figure 1.6. Intracellular LDLR pathway showing the binding and internalisation of APOB-LDLR complexes. The influence of LDLRAP1 and PCSK9 in the process is also represented. Created with BioRender.com.

Two thirds of the LDL particles are usually removed through LDLR, whereas the remaining is uptake by scavenger cell receptors of extrahepatic tissues [22]. Scavenger receptors of macrophages are responsible for recognition of LDL that was modified, mainly oxidised by free radicals, which represents the primary lesion of atherosclerosis, as explained below [5], [6].

1.5.3.4. Reverse cholesterol transport

ApoA-I is secreted as lipid-free protein from intestine and liver, it interacts with ATP binding cassette subfamily A member 1 (ABCA1) on basolateral membranes of enterocytes, hepatocytes, and macrophages, then attaining free cholesterol and phospholipids to form a

stable nascent HDL particle (Figure 1.7). Transition of disc-shaped nascent HDL to spherical mature HDL requires esterification of cholesterol by LCAT and its cofactor apoA-I, contributing to the formation of HDL hydrophobic core that is composed of cholesteryl esters. The subsequent addition of cholesterol to HDL occurs in macrophages and other peripheral cells through ATP binding cassette subfamily G member 1 (ABCG1) and scavenger receptor B-1 (SRB1) – molecules that prefer larger HDL as acceptors. In other hand, cholesteryl esters are transferred from HDL core to TG-rich lipoproteins in exchange of TG, a reaction catalysed by CETP (Figure 1.7). When depleted of cholesteryl esters, TG-enriched HDL is hydrolysed by HL with production of a smaller HDL particle that appears to be more avidly removed in kidneys. In addition, the phospholipid transfer protein (PLTP), structurally similar to CETP, catalyses the transfer of unsaturated FA present on phospholipids of apoB-containing lipoproteins to HDL [5], [8].

Cholesteryl esters within spherical HDL are transported back to liver by two mechanisms (Figure 1.7): transfer by CETP from HDL to apoB-containing lipoproteins, which are taken up through LDLR or LRP; direct delivery to liver through SRB1 with help from its adapter protein, the PDZ-domain-containing protein (PDZK1). Free cholesterol is excreted directly into bile or converted into bile acids by 7αH. Both processes result in delivery of sterol from peripheral tissues through plasma into hepatocytes, promoting the excretion of sterols into the stool [5], [8].

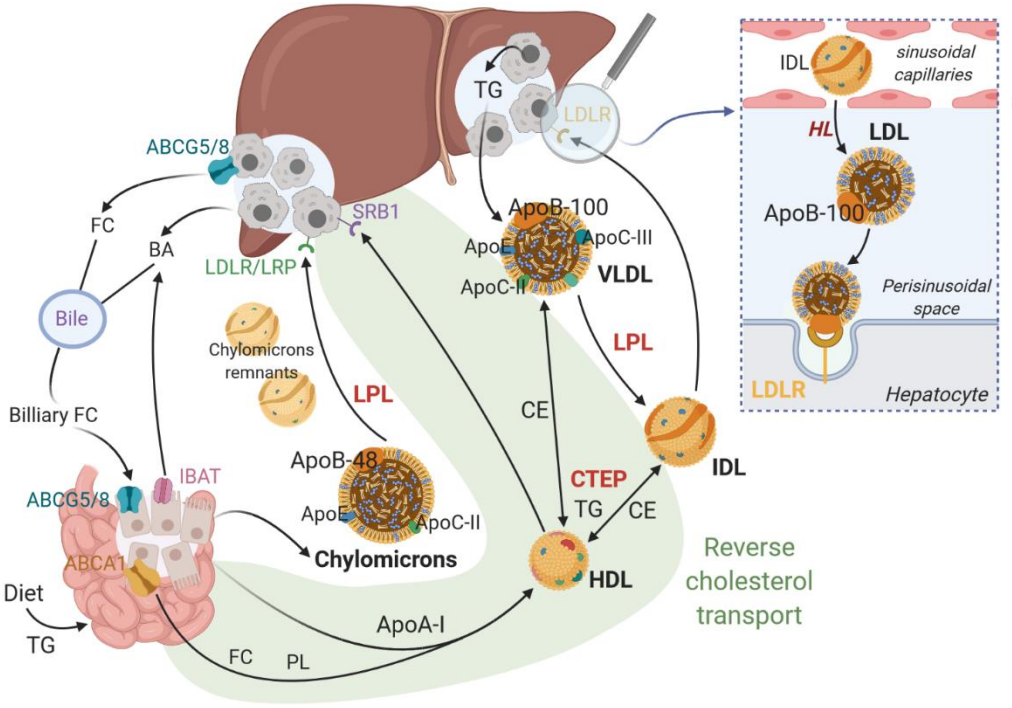


Figure 1.7. Reverse cholesterol transport and its connections to the remainder lipoprotein metabolism pathways. BA: bile acids; FC: free cholesterol; FFA: free fatty acids; CE: cholesteryl esters; TG: triglycerides; CETP: cholesteryl ester transfer protein; LPL: lipoprotein lipase; HL: hepatic lipase; PL: phospholipids. Created with BioRender.com.

Reverse cholesterol transport explains in part the protective effect of HDL and apoA-I against atherosclerosis and cardiovascular disease. Indeed, factors inhibiting this process, such as dysfunctional HDL, appear to promote atherosclerosis. In addition to its central role in reverse cholesterol transport, HDL may be cardioprotective through its antioxidant, anti-inflammatory and antithrombotic effects. For example, HDL inhibits LDL oxidation by metal ions, which may be due to the influence of several molecules on HDL, including apoA-I, platelet activating factor acetylhydrolase and paraoxonase. Accumulation of large HDL_2, which appears to be the most cardioprotective of HDL subclasses, is favoured by oestrogens that negatively regulate HL. Conversely, progesterone and androgens, which positively regulate HL, lead to increased production of small HDL_3 [5], [8]. Together with CETP, HL is believed to reduce the core of large HDL_2 particles and play a role in reconversion of HDL_2 to HDL_3 [14].

1.6. Cholesterol homeostasis

The cholesterol balance results from the interplay between cholesterol synthesis, absorption and excretion through bile and stool. When cholesterol excretion rises or its absorption diminish, cholesterologenesis is higher, whereas a higher content of cholesterol in tissues leads to inhibition of its endogenous synthesis. Indeed, as intracellular cholesterol decreases, LDLR activity increases and vice versa [23], [24]. This regulated feedback mechanism begins with the SREBP cleavage activating protein (SCAP), which is both a sensor of sterols and a chaperon of SREBP. When hepatocytes are deprived of cholesterol, SCAP transports SREBP from endoplasmic reticulum to Golgi apparatus, where site-1 and site-2 proteases release the NH2-terminal of SREBP from membrane to nucleus. There, it binds to a sterol response element (SRE) on promoter of LDLR and HMG-CoA reductase genes, thus increasing their transcription, which allows LDLR increase and subsequently low-density lipoprotein cholesterol (LDL-C) decrease. On the other hand, SREBP cannot reach the nucleus when hepatic cholesterol is increased, which decreases transcription of LDLR and HMG-CoA reductase genes, consequently decreasing LDLR and increasing LDL-C [5], [25]. In addition, when free cholesterol is released within lysosomes, there is an increment in production of cholesteryl esters by ACAT [24].

Lysosomal acid lipase (LAL), a 46 kDa glycoprotein encoded by *LAL* gene (also known as *LIPA*), is responsible for hydrolysis of lipoprotein cholesteryl esters and TG. Hydrolysis of cholesteryl esters generates free cholesterol, which after leaving the lysosome can be re-esterified in endoplasmic reticulum to form cytosolic lipid droplets. If excess free cholesterol

is retained in the lysosome, it can inhibit both lysosomal and LAL activities, which contributes further to progression of atherosclerosis [26].

2. Dyslipidaemia

The term dyslipidaemia corresponds to the presence of increased or decreased levels of lipoproteins in circulation [13]. Dyslipidaemia is one of the major cardiovascular risk factors especially when it is associated with increased levels of serum LDL-C and/or reduced levels of HDL-C [6], [27]. When serum LDL exceeds a threshold concentration, LDL particles traverse the endothelial wall and can become trapped in the arterial intimae, where they may undergo oxidation or other biochemical modification. These modified LDL particles can be taken up by macrophages, thus stimulating atherogenesis (Figure 1.8), which is the pathological process that leads to atherosclerosis – a disease of large and intermediate-sized arteries in which fatty lesions (i.e., atheromatous plaques) develop on arterial lumen [4], [6]. As a silent condition, dyslipidaemia does not usually produce clinical manifestations until an atherosclerotic vascular event occurs (e.g., myocardial infarction, stroke, or peripheral vascular occlusion) [6].

2.1. The development of atherosclerosis and the atherogenic role of LDL

The atherosclerotic process begins with the appearance of an initial vascular lesion that increases the expression of adhesion molecules on endothelial cells, while decreasing the ability of these cells to release nitric oxide and other substances that help to prevent adhesion of macromolecules, platelets, and monocytes to the endothelium. Once this initial damage occurs, circulating monocytes and LDL begin to accumulate at the site of injury [4]. As mentioned before, excess LDL that become trapped in arterial intimae may undergo biochemical modifications (e.g., oxidation or glycation). Monocytes also enter in the arterial intimae and differentiate in macrophages. In its turn, the modified LDL binds to macrophage scavenger receptors (Figure 1.8), like scavenger receptor A-1 (SRA1) and cluster of differentiation 36 (CD36) and is taken up by macrophages through a low-affinity and LDLR-independent mechanism, which is not subject to the feedback inhibition of LDLR synthesis by LDL-derived cholesterol. This gives macrophages a foam-like appearance, so that they are called foam cells, which aggregate and form a visible fatty streak on the blood vessel [4]–[6]. Then, foam cells trigger inflammation and growth of both the intimal layer and atherosclerotic plaque, while increasing susceptibility for the development of more lesions. With time fatty streaks grow larger and coalesce, the surrounding fibrous and smooth muscle tissues proliferate to form even

larger plaques. In the last instance, this can lead to occlusion or rupture of the arterial vessel (Figure 1.8) [4].

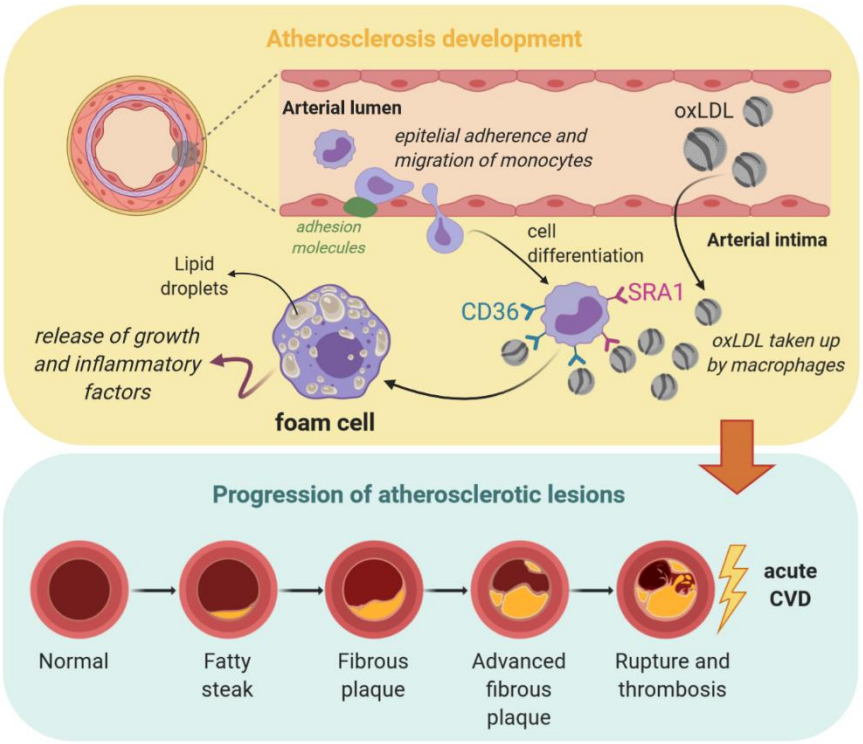


Figure 1.8. Development of an atherosclerotic plaque with production of foam cells (yellow shade) and the progression of atherosclerotic lesions, showing the different degrees of severity until the occurrence of an acute event (blue shade). Created with BioRender.com.

LDL is considered the most atherogenic lipoprotein, as well as the one with the greatest cholesterol proportion. LDL plasma population is composed of heterogeneous subfractions that are different in size, density, charge, and protein/lipid content. These LDL subfractions are generated during delipidation of VLDL to IDL and LDL particles. There are several techniques that allow the partition and identification of the different LDL subfractions, including nuclear magnetic resonance, high-performance liquid chromatography, ultracentrifugation, dynamic light scattering and gel gradient electrophoresis (GGE) [28], [29]. In GGE, a semi-quantitative method that detects predominance of smaller particles among LDL subclasses, two distinct electrophoresis-based phenotypes with peak LDL particle diameters >25.5 nm and ≤25.5 nm are usually determined, corresponding to phenotype A and B respectively [28]. Lipoprint™ is an example of semiautomated ready-to-use system for GGE, either specific to LDL or HDL particle profiling [30]. In this system, phenotype A is characterised by predominance of large buoyant LDL (lbLDL) and phenotype B is characterised by predominance of small dense LDL (sdLDL) [28], [29]. It has been established that sdLDL is the most atherogenic LDL subfraction, being already recognized as a cardiovascular risk factor. Indeed, subjects with phenotype B

have an increased risk of myocardial infarction and often exhibit atherogenic lipid profiles, such as high TG and low HDL-C levels [28].

Gender and menopausal status have significant impact on small dense low-density lipoprotein cholesterol (sdLDL-C) and sdLDL-apoB concentrations. Furthermore, the protein/lipid composition of LDL is strongly influenced by changes in TG concentration [29]. It has been proposed that the relative amounts of VLDL and LDL secreted into plasma reflect the body status of TG and cholesterol pools that need to be mobilised during metabolism [31]. The atherogenic potential of sdLDL can be explained by the following metabolic features:

1. small size that allows an easier infiltration on arterial wall;
2. high affinity with proteoglycans in arterial wall, allowing sdLDL to stay longer in the subendothelial space;
3. lower affinity with LDL receptors rather than larger LDL, which means a prolonged half-life in plasma, suggesting an impaired clearance from circulation;
4. lack of vitamin E that makes it highly susceptible to oxidation [28], [32].

In general, sdLDL levels increase in hypertriglyceridaemia, familial combined hyperlipidaemia, metabolic syndrome, and coronary artery disease. Nonetheless, it has been shown that atorvastatin can decrease sdLDL-C to an equivalent level of control individuals [28], [32].

2.2. Classification of dyslipidaemias

In addition to nutritional and secondary causes (e.g., obesity, diabetes mellitus or hypothyroidism), dyslipidaemia can occur as a consequence of specific genetic contexts [33]. In that case we are in the presence of a primary dyslipidaemia, which can be caused by a single gene variant (monogenic dyslipidaemia) or result from the cumulative effect of several polygenic and environmental factors (polygenic dyslipidaemia) [34]–[36].

The most important lipid disorders affecting cardiovascular risk are those that increase LDL or reduce HDL levels. Patients with monogenic dyslipidaemia are usually the most severely affected individuals. However, the majority of hyperlipidaemic patients do not have a monogenic defect, rather they are most likely to have a polygenic disease with a variable contribution from environmental factors (e.g., excessive saturated fat intake for high LDL levels, obesity or smoking for lower HDL levels) [6], [37]. Thus, it is important to distinguish between monogenic and polygenic dyslipidaemia, since monogenic patients are exposed to high cholesterol levels from birth and therefore have an increased risk of premature cardiovascular

disease. In these cases, prompt and accurate diagnosis is essential for cardiovascular disease prevention as it allows earlier and/or more aggressive therapeutic measures, which have been shown to be effective in reduction of cardiovascular morbidity and mortality in both adults and children [38].

At the moment, there is no well-defined and consensual classification system for dyslipidaemias. In 1965, Fredrickson and Lees grouped the hyperlipoproteinemias into five different types, each corresponding to a specific lipoprotein profile, independently of being primary or secondary hyperlipidaemias [39]. This classification system was formally adopted by WHO in 1970 [40] and during decades had an invaluable role for disease definition [9]. However, this classification is outdated for the following reasons: 1) only hyperlipidaemias are included; 2) absence of any condition primarily characterised by HDL-C deviations; 3) it is based solely on phenotype that is known not to be genetically specific, since it can reflect in a similar way monogenic, polygenic, or even secondary dyslipidaemias [9], [33]. Therefore, the present work uses a specific classification system based on the current knowledge regarding dyslipidaemias, as described in the following sections.

2.2.1. Primary monogenic dyslipidaemias

Based on the current literature, primary dyslipidaemias can be divided in five main groups: hypertriglyceridaemias, hypercholesterolaemias, mixed hyperlipidaemias, primary alterations of HDL, and hypolipidaemias associated with deficiency in apoB-containing lipoproteins [5], [8], [9]. Table 1.3 summarises the main monogenic dyslipidaemias following this classification and provides information regarding heritability pattern, associated genes and phenotype. Giving the scope of this thesis, the present text will focus on Familial Hypercholesterolaemia, as explained next.

Table 1.3. Characterization of the main monogenic dyslipidaemias according to heritability pattern, affected genes and phenotypic traits [5], [8], [33], [41]–[43].

Condition		Heritability	Related genes	Phenotypic traits
Hypertriglyceridaemias	Familial chylomicronaemia	Autosomal recessive	LPL APOC2 GPIHBPI APOA5 LMF1	Extremely high levels of TG and TC, pancreatitis, xanthomas, *lipaemia retinalis*, hepatosplenomegaly
Hypercholesterolemias	Familial hypercholesterolaemia	Autosomal co-dominant	LDLR APOB PCSK9	High levels of LDL-C, presence of tendinous xanthomas and premature coronary disease
	Autosomal recessive hypercholesterolaemia	Autosomal recessive	LDLRAP1	Very high levels of TC, xanthomas, premature CVD
	Sitosterolaemia (or phytosterolaemia)	Autosomal recessive	ABCG5 ABCG8	Extremely high levels of dietary vegetal sterols, tendinous xanthomas, premature CVD, haemolytic anaemia and macrothrombocytopenia
	Deficiency of 7αH	Autosomal co-dominant	CYP7A1	High levels of TC and sometimes of TG, biliary lithiasis
Mixed hyperlipidaemias	Familial dysbetalipoproteinaemia	Autosomal recessive (incomplete/low penetrance)	APOE (ε2/2)	High levels of VLDL remnants, IDL and chylomicrons, occurrence of an abnormal lipoprotein (β-VLDL), xanthomas, corneal arcus and xanthelasma
	Familial deficiency of hepatic lipase	Autosomal recessive	LIPC	High levels of TC and TG, xanthomas, and corneal arcus
Alterations of HDL	Familial hypoalphalipoproteinemia	Autosomal dominant	APOA1	Low HDL-C levels (<5th percentile, age and sex specific)
	Familial deficiency of LCAT	Autosomal recessive	LCAT	High levels of TC and free cholesterol, low HDL-C levels, visual perturbations, haemolytic anaemia, and severe nephropathy
	Tangier disease	Autosomal recessive	ABCA1	Extremely low HDL-C levels (<10 mg/dl), premature coronary disease and haematological alterations
Hypolipodaemias		Autosomal dominant/ recessive	APOB PCSK9 ANGPTL3 MTTP SAR1B	Marked reduction of LDL-C and TC levels

2.2.1.1.Familial Hypercholesterolaemia

Familial Hypercholesterolaemia (FH), an autosomal co-dominant disorder (i.e., both variant alleles from each affected parent contribute to the phenotype), is the most common monogenic dyslipidaemia, with an estimated prevalence of 1/250 for heterozygotes worldwide [33], [44].

The prevalence of homozygous FH is approximately 1/300 000. Still, heterozygous FH can be more frequent (up to 1%) in certain founder populations, owing to a high prevalence of embedded pathogenic variants [33]. This condition usually results from loss-of-function genetic variants in the *LDLR* gene and, less commonly, in the *APOB* gene. Gain-of-function variants in *PCSK9* gene also produce the same phenotype, although it is a rare cause of FH. Based on the inheritance pattern of co-dominance, FH patients can be considered as simple homozygotes, when the same variant is present in both alleles of the same gene; compound heterozygotes, with different variants in each allele of the same gene; or double heterozygotes, rare cases of individuals presenting variants in two different genes. In any of these scenarios, the offspring will be obligately heterozygote, assuming that the other parent has not FH [33], [45].

This disorder is clinically characterised by high LDL-C levels since birth and consequent accumulation of cholesterol in peripheral tissues (occurrence of tendinous xanthomas and corneal arcus) and arteries, which triggers the premature development of atherosclerosis and increases the risk for coronary heart disease. In comparison to the normal population, FH patients have a 100-fold increased risk to develop premature cardiovascular disorders (i.e., at 20-39 years old). FH is still widely under-diagnosed and undertreated, including in Portugal, where no prior clinical or genetic studies have been performed until the beginning of the Portuguese FH study (PFHS) in 1999 [46], [47]. An early diagnosis of FH is essential to improve the prognosis by administering appropriate therapeutic measures, genetic counselling, and access to specialised medical services [35], [47].

2.2.1.1.1. Molecular studies in the hallmark FH genes

The PFHS has been performing a systematic characterisation of FH cases in Portugal and includes extended lipid analysis profiles for a large number of index patients, as well as molecular studies involving the known FH-related genes [48], [49]. These molecular studies are performed in five phases as follows:

1) study of the promoter, splicing and coding regions of the *LDLR* gene and screening for the most common *APOB* variants (fragments of exons 26 and 29);
2) study of large rearrangements by Multiplex Ligation-dependent Probe Amplification (MLPA) technique for *LDLR*;
3) study of the promoter, splicing and coding regions of the *PCSK9* gene;
4) study of promotor, all exons, and flanking regions of *APOB*;
5) functional *in vitro* studies, only in certain cases as explained forward [50], [51].

Phases 3 and 4 are performed only if no putative variants are found in the first two phases. In phase 3, the study of the whole *PCSK9* gene is performed for severely affected patients [51]. All reported variants are annotated according to the Human Genome Variation Society. The variants are considered "novel" when they are not present in the University College London LDLR FH database or in the Human Gene Mutation Database at the time of the molecular studies [50], [51].

LDLR variants can be grouped in two categories: receptor-negative or null variants that result in either no protein synthesis or synthesis of a completely non-functional protein, and receptor-defective (or simply defective) variants that result in synthesis of an ineffective protein [10], [33]. Accordingly, null variants keep up to 2% of the molecular activity from wild type allele, while defective variants contribute for the maintenance of 2-80% of the molecular activity [52]. *LDLR* null variants comprise large-scale copy number variations (usually partial gene deletions), nonsense variants in the coding region, and splicing variants that are typically non-coding variants and occur at intron-exon boundaries. These variants are usually found in individuals with the highest LDL-C mean values. Conversely, defective variants are associated with lower LDL-C mean levels and they include missense variants, which correspond to the alteration of a single amino acid residue, and small insertions or deletions (indels) within or near the coding sequence, some of which might shift the reading frame (frameshift variants) [10], [33]. Although *LDLR* variants affect the protein in variable ways and conduct to different degrees of disease severity, all kinds of *LDLR* variants lead to impaired uptake of circulating LDL-C and consequent rise of its serum levels [53].

Attributing FH causality to *APOB* and *PCSK9* variants is more complicated in comparison to variants found in *LDLR* gene, since these two genes are very polymorphic and *in vitro* functional studies are more difficult to perform. Frameshift, missense, nonsense, and splicing variants were already found in the *APOB* gene [33]. Still, *APOB* variants do not have a penetrance of 100% and the resultant phenotype is usually milder in comparison to that produced by *LDLR* variants [50]. Most common *APOB* variants are present in exons 26 and 29, which correspond to the coding region of the receptor-binding domain of apoB protein that is essential for LDLR-apoB binding and, consequently, for circulating LDL-C uptake. In the case of *PCSK9*, most of the reported variants are missense variants, although frameshift variants have already been found [33], [54]. As PCSK9 protein is responsible for LDLR degradation in liver cells, gain-of-function variants result in increased *PCSK9* activity that leads to an increased intracellular LDLR degradation, allowing LDL-C to accumulate in circulation [53].

The assessment of the functional effect of identified genetic variants is one of the main challenges in FH molecular diagnosis. A study of 2015 reported that about 30% of detected variants in the PFHS cohort had an uncertain pathogenic effect [49]. According to the American College of Medical Genetics and Genomics (ACMG) guidelines, variants are classified as pathogenic, likely pathogenic, variant of unknown significance, likely benign, and benign. If a variant has been described in more than 5% of the studied population - i.e., minor allele frequency (MAF) >5%, in the 1000 Genomes database or the Exome Sequencing Project database, it is considered a common variant or a polymorphism and thus a neutral variant [51]. In a study from 2018, reporting the result of functional studies of *APOB* variants from individuals of PFHS, a MAF >1% was considered to define a common variant [50]. For variants without functional studies, several software tools are applied for *in silico* analysis, namely prediction of amino acid substitutions and splicing defects. In PFHS, functional studies are only pursued for missense variants, in-frame deletions/insertions, and splicing variants when no functional assays have been performed yet and external funding is available [51].

All functional studies performed in the context of PFHS comprise *in vitro* assays and the process is mostly similar for all studied genes, although with some specificities for *APOB* and *PCSK9* [50], [51], [55]. Given the fact that the majority of FH-causative variants are identified in the *LDLR* gene, functional studies are here briefly explained for LDLR variants. The main purpose of analysing these variants is to measure their ability to interfere with protein expression, binding and internalisation. The *in vitro* assays use cell lines transfected with the *LDLR* variant to be studied, which are kept in cell culture for 48h to achieve the maximal *LDLR* expression. Western blot analysis evaluates the *LDLR* expression. The quantification of protein expression is performed using flow cytometry, which determines the expression of the receptor at the cell surface while comparing the fluorescence produced by the variant in study with the fluorescent signal from wild type *LDLR* expressing cells (controls). The same technique of flow cytometry, namely fluorescence-activated cell sorting (FACS), is used to quantify *LDLR* activity, with determination of LDLR-LDL binding and LDL uptake. The LDL particles used in this step are isolated from serum samples of healthy individuals. Confocal laser scanning microscopy is used to analyse *LDLR* expression and intracellular colocalization. In addition, the kinetics of *LDLR* variants are evaluated by measuring the protein expression at different incubation times in the presence of LDL particles, as well as by performing studies of LDLR-LDL binding at different pH values [49], [56]. Although the cut-off value for determining whether an *LDLR* variant is considered a functional mutant has not been established yet, several studies have been considered that a value of *in vitro LDLR* activity in the affected allele below

70-80% of wild type can classify a variant as pathogenic. This value corresponds to 85-90% of total *LDLR* activity in both gene alleles, assuming that the non-affected allele can produce 50% of the active protein. The establishment of functionality cut-offs contributes for a proper classification of gene variants, which is an important step in the direction of a personalised treatment for FH. This requires more functional studies and an integrated analysis of clinical, molecular and functional data, allowing an accurate FH diagnosis [49]. A detailed analysis of worldwide submitted FH-related variants in ClinVar, an NCBI-funded resource, is provided by Iacocca et al. on behalf of the ClinGen FH Variant Curation Expert Panel [57]. This is essential to achieve a standardised and reliable variant classification, allowing an improvement of FH diagnosis.

2.2.1.1.2. Clinical criteria

Two clinical scoring systems are in general use for FH diagnosis, the Simon Broome (SB) criteria and the Dutch Lipid Clinic Network criteria [33], [58]. Other proposed systems include the one used by the American Heart Association, and the Canadian simplified FH definition. Although the concordance between the different algorithms is inconsistent, most of them score and assign weights to the following features:

- lipid values, particularly total cholesterol and/or LDL-C;
- presence of physical stigmata considered pathognomonic for FH, such as tendon xanthomas, xanthelasmas, or *arcus cornealis*;
- personal or family history of premature cardiovascular disease;
- pathogenic or likely pathogenic DNA variants in FH-related genes [33].

Physical stigmata were observed in more than half of FH patients in the 1970's, but nowadays they are reported only in 5-20% of well-characterised FH cohorts, which might be due to an earlier diagnosis and treatment - with prompt adoption of a healthy lifestyle and drug therapy, mostly based on statins [33].

Since the work developed during this thesis involved the application of SB criteria, the same are presented in detail in Table 1.4.

Table 1.4. Simon Broome clinical criteria applied for FH diagnosis [33], [59], [60].

Confirmed/Definite FH
Biochemical Profile - age < 16 years: TC ≥ 260 mg/dL or LDL-C ≥ 155 mg/dL - age ≥ 16 years: TC ≥ 290 mg/dL or LDL-C ≥ 190 mg/dL AND *Physical stigmata* - tendinous xanthomas at index case or relatives (parents, sons, grandparents, siblings, uncles) OR *Genetics* - evidence of a pathogenic or likely pathogenic variant in *LDLR*, *APOB* or *PCSK9* genes
Probable FH
Biochemical Profile - age < 16 years: TC ≥ 260 mg/dL or LDL-C ≥ 155 mg/dL - age ≥ 16 years: TC ≥ 290 mg/dL or LDL-C ≥ 190 mg/dL AND *Family history* - myocardial infarction: before 50 years in grandparents and uncles, or before 60 years in parents, siblings, or sons OR - LDL-C ≥ 290 mg/dL in parents, siblings, sons, grandparents, or uncles

Current guidelines support genetic testing of the hallmark FH genes (i.e., *LDLR*, *APOB*, *PCSK9*) for individuals that comply with clinical diagnostic criteria (biochemical profile and family history of cardiovascular disease) [61]. When genetic evidence of FH is found, genetic testing should be extended to patient relatives, following the system of familiar cascade screening to identify new cases of FH [62].

2.2.1.1.3. Drug treatment

The first step of FH treatment is the adoption of a healthy lifestyle, including a diet with reduced fat intake, physical exercise, and avoiding harmful habits like smoking. Still, this is usually not enough to manage cholesterol levels and pharmacotherapy is the norm [10], [63]. Statins, which are HMG-CoA reductase inhibitors, are the most commonly used drug to reduce cholesterol blood levels, being considered first-line drug treatment for FH management, including for children. When statins are not enough to achieve cholesterol goals, they are often used in combination with ezetimibe, the only available cholesterol absorption inhibitor, which blocks the intestinal Niemann-Pick C1-like protein 1 (NPC1L1) that mediates cholesterol uptake by enterocytes [10], [54], [63]. Ezetimibe results in a selective lowering of cholesterol absorption without the impairment on absorption of other nutrients [64]. Several studies have shown that adding ezetimibe to any dose of a statin leads to a 20% additional reduction in LDL-C [10].

Additionally, other lipid-lowering drugs are available for dyslipidaemic patients, as seen in Table 1.5.

Table 1.5. Lipid-lowering drugs currently available for the treatment of FH [10], [54].

Drug class	Mode of action
Statins	Inhibition of HMG-CoA reductase
Ezetimibe	Cholesterol absorption inhibitor
Bile acid sequestrants/Resins	Sequestrate micelles promoting their excretion before being captured by enterocytes
Niacin	Reduce FFA mobilisation from adipose tissue to the liver, impairing lipoproteins synthesis
Human monoclonal PCSK9 antibodies	Inhibition of LDLR-PCSK9 binding with promotion of LDLR recycling and LDL clearance
Lomitapide	MTTP inhibitor reducing the assembly and synthesis of lipoproteins both in enterocytes and hepatocytes
Mipomersen	Antisense oligonucleotide binding apoB mRNA with reduction of VLDL and LDL generation

Statins can be classified according to source (natural or synthetic), hydrophilicity (hydrophilic or hydrophobic) and lipid-lowering power. Hydrophilic statins are almost or completely independent of the CYP450 system, being excreted mostly unchanged and less subject to pharmacokinetic interactions. In addition, these statins are more liver-specific by using active transporters to be taken up by hepatocytes, compared with hydrophobic statins that can passively diffuse through cell membranes [54]. This translates in a lower risk of side effects, although high efficacy and safety has been demonstrated for all statins in both adults and children [54], [65]. Still, statin-associated muscle pain or weakness may occur, especially after a long-term high-dose treatment. In this case, water-soluble statins are preferred (e.g., pravastatin or rosuvastatin). Regarding the lipid-lowering power, the most potent statins are rosuvastatin, atorvastatin and simvastatin [54].

For severe cases of FH, like homozygous FH patients, lipoprotein apheresis, a extracorporeal procedure similar to dialysis that selectively remove LDL particles from blood, can be applied in combination with a drug treatment plan [10], [63].

2.2.2. Polygenic dyslipidaemia

Functional polymorphisms in the hallmark FH genes (*LDLR*, *APOB*, *PCSK9*), or in other genes involved in lipid metabolism (e.g., *APOE* and *SREBP*), have been associated with minor functional changes in the encoded proteins which, although responsible for a slight increase in cholesterol levels, do not cause disease by itself. Still, the presence of these polymorphisms in multiple genes can have a cumulative effect, increasing the overall pathogenic outcome, thus leading to a polygenic form of dyslipidaemia [10], [33]. Conversely, some LDL-C lowering

variants and polymorphisms have already been identified in the three known FH-related genes and in *ANGPTL3* (loss-of-function variants). This may contribute to reducing the pathogenic outcome that results from LDL-raising single nucleotide polymorphisms (SNPs) or attenuate the effects of FH-causative variants in monogenic patients [33], [66].

A study by Talmud et al. proposed a weighted LDL-C genetic risk score that comprises 12 common LDL-C associated SNPs from genome-wide association studies (GWAS) - see below (section 3.3.1), which constitute LDL-C rising alleles and were identified through the Global Lipids Genetics Consortium. This kind of genetic scores can be useful to identify a potential polygenic contribution in individuals that fulfil FH clinical criteria but fail to present a pathogenic gene variant in any of the FH-associated genes, thus allowing the assessment of cardiovascular risk in these patients. Accordingly, polygenic individuals have a significantly higher mean LDL-C score than the general population [36]. Although inheritance of polygenic dyslipidaemia is not well-defined as FH, which follows an autosomal co-dominant Mendelian inheritance pattern, the screening of family members is recommended since the LDL-C raising alleles cluster in families [33].

Several studies have been shown that although polygenic risk scores may be useful as markers of hypercholesterolaemia severity and help predicting CVD risk, they do not seem to be a reliable tool for discrimination between monogenic and polygenic individuals [66]–[68]. A polygenic score might be used to detect the genetic cause of hypercholesterolaemia, after failing the detection of a large-effect variant in any of the FH hallmark genes, both in adults and children [67], [69].

2.2.3. Other genetic and non-genetic contributions to dyslipidaemia

The presence of a pathogenic or likely pathogenic variant in any of the FH-associated genes may explain the high LDL-C levels only in a part of the suspected FH individuals [33]. Other genetic causes could be beyond the hypercholesterolemia in patients that fail to confirm a diagnosis of FH in molecular studies, including the identification of pathogenic variants associated with FH phenocopies (e.g., variants in *ABCG5/ABCG8*, *APOE* or *LIPA* genes), autosomal recessive hypercholesterolaemia (*LDLRAP1* variants), or a polygenic effect (as explained previously). Genetically-driven high levels of Lp(a) can also raise LDL-C levels and increase the risk of CVD, which can explain 5-20% of suspected FH cases [10], [33]. Indeed, the European atherosclerosis society (EAS) recommends the measurement of Lp(a) concentration in FH patients, since a substantial proportion of these individuals have a lifelong

elevation of Lp(a), which is considered higher when it raises above 50 mg/dL [70]. High LDL-C levels can also be caused by non-Mendelian mechanisms like environmentally induced epigenetic effects [33]. In fact, non-genetic factors are known to constantly modulate the final phenotypic expression, including effects of diet and lifestyle [10], [33]. Therefore, the possible combination of multiple genetic and non-genetic factors might explain the heterogeneity registered among individuals with suspected FH, even in patients with confirmed diagnosis and presenting the same FH-causative variant [10], [33], [45].

In the past two decades, a tendency to develop an increasingly unhealthy cardiovascular risk profile has been reported in individuals of age 18-50 years, especially regarding an increased prevalence of overweight/obesity. The incidence of CVD in young adults has mostly been steady or slightly increasing, in contrast to the generally decreasing trends observed in older individuals. This may be explained by the presence of multiple CVD risk factors associated with an unhealthy lifestyle (e.g., physical inactivity, poor diet, smoking, or obesity) and consequent conditions like diabetes, hypertension, or dyslipidaemia. These observations suggest a future increase in the cardiovascular burden as the younger individuals age [71].

3. Biomarker discovery in the context of dyslipidaemia

3.1. New biomarkers are needed to improve the selection for genetic screening

As mentioned before, given the silent nature and prevalence of FH, current guidelines support the testing of *LDLR*, *APOB* and *PCSK9* genes in patients that comply with clinical diagnostic criteria, and cascade screening of their family members [61]. However, most hyperlipidaemic subjects do not have a monogenic defect [33], [72]. Rather, their disease is most likely established through a polygenic genetic background, with a variable environmental contribution modulating the phenotypic expression [33], [72]. Although the lipid profile of polygenic subjects is usually less severe than that of FH subjects regarding total cholesterol (TC) and LDL-C levels, the differences are often subtle enough to prevent an accurate distinction between the two [54]. As a consequence, the yield of FH genetic screening programs is relatively low, assuming significant costs for patients and/or national health systems.

The analysis of data from children is particularly important, since it is well known that the correct identification and stratification of individuals at major cardiovascular risk during childhood allows an early implementation of a healthy lifestyle, and lipid lowering therapy when needed, thus reducing the cardiovascular burden in general population [35], [63].

Furthermore, children are not regularly submitted to blood tests for lipid profiling and their normal values may differ substantially from those of adults [60].

Previous work using PFHS data revealed that approximately 60% of the children that complied with SB criteria were negative for pathogenic or likely pathogenic variants in the hallmark genes, most likely corresponding to cases of polygenic/environmental hypercholesterolaemia [35]. These individuals will be referred to as FH-negative (FH-) along the text, while those that both fulfil SB criteria and present pathogenic or likely pathogenic variants in one of the FH-related genes will be referred to as FH-positive (FH+).

In the same study, FH+ subjects showed higher concentration of atherogenic particles (LDL-C) and lower concentration of anti-atherogenic particles (HDL-C), contrary to FH- individuals that presented higher levels of TG, apoC-II, apoC-III, apoE, as well as higher frequency of overweight/obesity [35]. This suggests that the integrated analysis of multiple biomarkers can be used to create a model that can effectively discriminate between these two populations, improving the selection of patients for genetic screening. Furthermore, a better understanding of the lipid profiles of FH+ and FH- patients may shed light on the molecular and genetic basis of polygenic hypercholesterolaemia, eventually leading to the identification of novel biomarkers and/or therapeutic targets [35], [63].

Adding to this, the standard methods of lipoprotein measurement fail to identify many lipoprotein abnormalities that contribute to cardiovascular disease risk. Advanced lipoprotein tests offer insight into subtle, yet relevant, aspects of lipoprotein metabolism and atherosclerosis that help to explain the eventual failure of LDL-C lowering strategies to stem atherosclerosis development. It has been suggested that apolipoprotein measurements could replace, or at least complement, the standard tests (measurement of total cholesterol, LDL-C, HDL-C and TG), since they are more accurate and reproducible, besides reflecting better the risk of coronary heart disease [12], [14]. Indeed, the measurement of apoB and apoA-I is already used as complementary to standard markers in the current guidelines for management of dyslipidaemia from the EAS [62]. Identification of novel biomarkers could thus contribute to the improvement of FH diagnosis and management by enhancing the selection of individuals for genetic testing and help with the assessment of CVD risk stratification.

3.2. Traditional approach for biomarkers in dyslipidaemia: establishment of cut-offs

One of the first steps for finding biomarkers is the analysis of data from a dyslipidaemic cohort comprising a group of clinical, biochemical, or molecular parameters, and compare the mean

or frequency value (if it is a continuous or categorical variable, respectively) of each parameter between FH+ and FH- individuals. Following a pair-wise approach, for each parameter statistical tests are performed to evaluate if the difference of means in the FH+ and FH- populations is statistically significant. Afterwards, considering as potential biomarkers the parameters with a significant difference between FH+ and FH- individuals, cut-offs are determined from receiver operating characteristic (ROC) curves that plots all the measurements of a given parameter regarding sensitivity and specificity. The value that maximises the sum of sensitivity and specificity, with sensitivity higher than specificity while both values are above 50%, is selected as the optimal cut-off point for each biomarker [35].

This approach corresponds to the search for a correlation between an individual parameter and the independent variable/outcome (subject's FH classification), which fails to represent the complex network of interactions between parameters that can occur under the biological context of dyslipidaemia. Although it is possible to perform additional analysis to measure the correlation between parameters, a traditional statistical approach like this does not allow us to get an integrative overview of the data and capture complex connections among parameters, as well as between them and the independent variable. This is a principal limitation of this kind of approach, which may explain the low specificity of some clinical criteria based on strict cut-offs, in comparison to a multiparameter approach (e.g., machine learning-based modelling) [35], [72].

3.3. Alternative methodologies in the search of biomarkers and integrative knowledge

New approaches are needed to achieve a better distinction between monogenic and polygenic dyslipidaemia patients, preferentially based on methods that support a more accurate and less time-consuming diagnosis, with an acceptable cost/benefit ratio. Application of integrative data analysis tools, such as machine learning-based modelling and clustering analysis, is expected to assist in the identification of sub-groups of individuals characterised by specific biological parameters, thus supporting the identification of a set of parameters that can best differentiate patients – potential biomarkers [73], [74]. Likewise, it is important to integrate this knowledge into the current understanding of lipid metabolism pathways and of the main genes involved, creating a solid base for identifying which genes/pathways can be affected (and how) in the context of polygenic dyslipidaemias. Therefore, the application of an integrative approach is essential for a competent analysis of such a complex biological question [75].

3.3.1. Advances in genetic profiling: GWAS and polygenic risk scores

Between 1990 and 2015, Sanger sequencing of polymerase chain reaction-amplified coding regions of *LDLR* and specific regions of *APOB*, and more recently of the *PCSK9* gene, was the most commonly used method for genetic diagnosis and detection of new variants associated to FH [76]. Despite being time consuming and expensive on a per-nucleotide basis, Sanger sequencing is still commonly used by diagnostic laboratories. Recently, next generation sequencing (NGS) techniques are starting to supersede Sanger sequencing using a variety of approaches that support massively parallel sequencing [77]. In contrast to Sanger sequencing, NGS can be easily applied to achieve whole-genome sequencing or whole-exome sequencing, generating data regarding the entire genome or only the coding DNA sequence respectively. Alternatively, NGS can be used within a targeted sequencing panel, designed to acquire information on a selected group of genetic regions known to be relevant to the disease of interest at a lower cost [78].

As mentioned before, monogenic dyslipidaemia, namely FH, increases the risk of developing premature cardiovascular disease. Still, most hyperlipidaemic patients do not present any pathogenic variant at the known FH genes. Instead, they are likely to suffer from a polygenic form of dyslipidaemia with a variable contribution from environmental factors (e.g., diet and lifestyle) [6], [35], [36]. Indeed, several GWAS have demonstrated that common DNA variants account for the majority of heritable risk for complex diseases like CVD [79].

The recent advances in whole-genome sequencing offer new tools for genetic testing and disease diagnosis, given the ability to capture the complete spectrum of genetic variation, both the rare single large-effect variants found in monogenic patients and the common variants of small effect, whose cumulative impact translates in polygenic dyslipidaemia. Polygenic scores quantify the genetic susceptibility conferred by the cumulative effect of multiple common variants into a single normally distributed risk factor [79].

In recent years, the analysis of several cohorts by GWAS has been carried out aiming to identify new variants and potential genes of interest associated with lipid traits. Such studies contribute to improving the diagnosis of dyslipidaemias and the cardiovascular risk stratification, as well as to a better integrated knowledge of lipid and lipoprotein metabolism in healthy and unbalanced states [80]–[82]. Furthermore, they may also contribute to the discovery of new drug targets and the assessment of the best treatment for each patient, opening doors for personalised medicine [83].

A study by Buscot et al. assessed the association between GWAS derived polygenic risk scores and LDL-C, HDL-C and TG trajectories from childhood to adulthood. Although the influence of genetic factors on age-specific lipoprotein values and developmental trajectories is complex, the developed and tested polygenic risk scores were shown to be highly predictive of LDL-C, HDL-C and TG levels at all the ages analysed [84].

In a work developed by Tabassum et al., GWAS was carried out on lipidomic profiles of 2181 individuals using about 9.3 million genetic markers, which was followed by a phenotypic-wide association study (PheWAS) that included 25 CVD-related phenotypes in more than half million of individuals. This study found that, similar to the common lipid traits (e.g., TC, LDL-C, HDL-C), there is a polygenic contribution to the abundance of lipid species (e.g., cholesteryl esters, TG, sphingomyelins), which may play a considerable role on the endogenous regulation of lipid metabolism. CVD risk was also associated with the lipidomic profile. The application of a GWAS approach on these lipid species was found to be useful in the identification of additional variants that could not be captured by traditional lipid and lipoprotein measurements [85].

Recently, Khera et al. developed and validated a polygenic score for early-onset myocardial infarction (EOMI) comprising a genome-wide set of 6.6 million common DNA variants (allele frequency above or equal to 1%), which has demonstrated a substantially better predictive capacity than a previous score restricted to 50 variants. Accordingly, a higher polygenic score was found among patients with EOMI in comparison with control subjects. Although monogenic and polygenic dyslipidaemia have a similar associated risk of EOMI, a high polygenic score is 10-fold more prevalent among EOMI afflicted individuals [79]. A polygenic score for coronary artery disease involving a similar number of common variants was developed in a previous study [81].

In comparison to the traditional risk assessment, these scores had the advantage that they can be assessed from the time of birth, well before the discriminative capacity emerges for risk factors like hypertension. In addition, making individuals with high polygenic scores aware of their inherited susceptibility may facilitate intensive prevention efforts [79], [81].

Notwithstanding the potential of polygenic scores for unrevealing the genetic causes of dyslipidaemia in patients without a monogenic cause, or for an improved disease prognosis and CVD risk stratification, there are still some outstanding issues to solve, such as the following: designation of a threshold to consider when the score is increased, proper integration of this score with other clinical and lifestyle factors, optimization of polygenic scores in individuals of non-European ancestry [79]. Moreover, there is still a considerable number of individuals with

hypercholesterolaemia that had neither a monogenic or polygenic explanation for their lipid levels, which suggests the contribution of unknown genetic or gene-environment causes [80]. At comparable levels of LDL-C, both monogenic FH and polygenic hypercholesterolaemia appeared to be associated with a considerable increased risk of CVD compared with hypercholesterolaemia with no identified genetic cause [80]. Nonetheless, the complex interplay between several genetic and environmental risks that lead to onset and progress of the condition are still poorly understood [84].

In the near future, it is expected that the potential clinical utility of polygenic risk scores will lead to a more generalised use of genome-wide analyses, also supported by the expected decrease in cost of NGS methods [86]. Recent studies have shown that the scores developed using a genome-wide approach are the most successful at assessing polygenic risk in complex diseases [79], [81].

3.3.2. Machine learning-based methods

Machine learning (ML) involves the development of statistical models and algorithms that can progressively learn from data and achieve desired performance on a specific task, especially when it is not possible to manually develop a set of rules based on all the intrinsic characteristics of the data. Within ML we can find both supervised and unsupervised learning methods, as present in the following subsections (3.3.2.1 and 3.3.2.2). Comparatively to traditional statistical methods, ML has better performance at finding patterns within complex datasets like, for example, clinical data. Thus, the application of ML based-methods can contribute to the identification of reliable disease biomarkers and improve patient stratification systems (e.g., disease classification, evaluation of cardiovascular risk or disease prognosis) [74]. Still, ML also has limitations, mainly being prone to bias and lack of interpretability from model classifiers, which means that it is not usually easy to understand how the model arrived at a given classification. This implies a difficulty in extracting relevant knowledge from the results of a model classification, especially concerning relationships between parameters contained in data or learned by the model [74], [87]. Regarding bias, this may occur when the distribution of the training data (i.e., part of a dataset used for training a ML-based model) does not reflect the characteristics of the real data that had served as source for modelling, which is known as the sample bias. Conversely, human bias might be present during the gathering and labelling of data used to train ML algorithms [74].

3.3.2.1. Supervised learning

Supervised learning requires a labelled training dataset, which means an *a priori* knowledge regarding the independent or target variable. Among supervised learning, regression or classification methods can be applied when the independent variable is continuous or categorical, respectively. Common supervised learning algorithms include linear regression, logistic regression, decision tree, support vector machines and artificial neural networks [74]. ML-based modelling relies on the computational capability of learning all the complex and non-linear interactions between variables, in a non-humanly achievable way, while keeping to a minimum the error between predicted and observed outcomes and thus improving prediction power [73].

A supervised ML approach may be prone to specific limitations including overfitting, highly correlated variables, unbalanced data (e.g., disproportional numbers of individuals between different classes in a dataset), or too small datasets that do not comprise a representative sample of the target population [73]. Overfitting happens when a model shows a high accuracy during training but then is not able to make robust predictions when applied to an independent dataset (unseen data). This problem is usually caused by the presence of noise (i.e., irrelevant information or randomness), which might be linked to a high number of variables present in the dataset versus the total number of observations [88], [89]. To detect overfitting, the dataset should be divided into a training and a testing set, so after training the model we can apply it to the testing set and evaluate how well the trained model will perform with new data. There are some options to avoid overfitting, including the use of cross-validation during training, increasing training dataset size, or applying methods for parameter selection that reduce the number of dependent variables. One of the methods for parameter selection is the exclusion of one element of a pair of highly correlated parameters, after measuring correlation of all possible pairs of parameters present in the dataset. Models comprising highly correlated parameters do not present additional information, while in turn these parameters may represent noise and interfere with model performance [89].

3.3.2.2. Unsupervised learning

Unsupervised learning aims to identify patterns within an unlabelled input dataset, requiring for example the application of methods based in dimensionality reduction (e.g., principal components analysis, PCA) or clustering analysis for identification of different groups/clusters of subjects among the dataset [74]. There are several clustering methods that can be categorised

according to the nature of data, criteria of the similarity measure, dimensionality, and scalability issues. Then, clustering methods can be mainly classified as hierarchical and non-hierarchical/partitional methods. Hierarchical clustering organises data into a hierarchical structure based on appropriate (dis)similarity or distance measures between every pair of subjects in the dataset. Several metrics can be used including Euclidean/Manhattan distances and correlation-based distances. In turn, there are two types of hierarchical clustering: agglomerative (or AGNES, from agglomerative nesting), when clustering groups data by means of a sequence of partitions starting with each unit (subject) forming a separate cluster and then merging similar clusters into larger clusters; divisive (or DIANA, from divisive analysis), when data clustering starts with one cluster comprising all units (subjects) and then splits it into consecutively smaller clusters. For partitional clustering, we should start by finding the optimal k (i.e., number of clusters) using any of several available methods, including the average silhouette, wss method (within cluster sums of squares), gap statistics, or simply apply the *NbClust* function in R (*NbClust* package) that combines different methods at once. Then, the clustering algorithm directly divides data units by the number of clusters previously determined, according to the distance metrics previously mentioned and following a non-hierarchical structure. As an example of partitional clustering, K-means is the most commonly used clustering method, where each cluster is represented by its centre (i.e. centroid) that corresponds to the mean of points (subjects) assigned to the cluster [90]–[92].

3.4. Understanding the biological context beyond dyslipidaemia: improving the lipid knowledge base

3.4.1. The concept of knowledge databases

Acquired data needs to be transformed into information, which in turn needs to be stored in an ordered way in (public) databases named knowledge databases (KDs) – or simply, knowledge bases. Thus, KDs present processed data extracted from experiments and/or computational assays [93]. They can be specific for different types of entities (e.g., genes, proteins, metabolites, or diseases), like Entrez Gene and UniProt, or even variations of the same topic (e.g., DNA sequence transcripts or SNPs), like dbSNP. Other types of KDs are not focused on individual terms but instead target physical interactions, such as REACTOME and KEGG, which consist of a collection of biological pathways/maps. Currently, KDs present some challenges, including interconnection between KDs to allow a better integrative view, public

access, updatability (able to include new data types), and definition of information storage standards [93].

The process of annotation, which comprises the assignment of properties to a given biological entity or the establishment of relations between those entities, is an essential step in the creation and updating of KDs. Annotation is based on evidence that can be classified as experimental, when inferred from direct assays, physical or genetic interactions, mutant phenotype, or gene expression patterns; or computational analysis evidence, when based on sequence or structural similarity, genomic context, or other analytical processes involving reviewed computational work [93]. Another important contributor for KDs is data integration, provided by network and visualisation tools (e.g., Cytoscape). This is one of the biggest challenges for present KDs, since the amount of produced data has been rising in recent years, mostly due to the development and application of high-throughput analysis systems, and it is generally considered in the scientific community that a single data type is not enough to offer a complete vision of any biological system [93].

3.4.2. The lack of data integration in lipid metabolism

Although most pathways of lipid metabolism are well known, a greater understanding is still required [75]. Lipids comprise up to a third of metabolomic database entries, but most of the online databases include a mix of curated and computationally generated lipids (e.g., LIPID MAPS, LipidHome, Human Metabolome Database), with the latter ones comprising only *in silico* "theoretical" lipids that may not exist in mammalian or biological systems (Table 1.6) [94].

Table 1.6. Examples of available databases regarding lipid species and its main structural and biological properties. A brief description with the main contents of these databases and their links are also shown [95], [96].

Database	Description
LipidHome	Database of theoretical lipids optimised for high throughput mass spectrometry lipidomics, which comprises an introduction of chemistry and biochemistry of individual lipid classes. http://www.lipidhome.co.uk/
LIPID MAPS (abbreviation of LIPID MAPS® Lipidomics Gateway)	Aiming to identify and quantitate, using a systems biology approach and mass spectrometry, all lipid species in mammalian cells, as well as quantitate the changes in the species in response to perturbation. It is associated with the LipidHome website. http://www.lipidmaps.org/
Lipid Bank	Contains diverse information regarding lipids, including molecular structures, nomenclature, spectral data (e.g., mass spectrometry, ultraviolet, infrared, nuclear magnetic resonance), and selected literature. http://lipidbank.jp/

On the other hand, there are some metabolic databases containing tissue-specific information, namely regarding liver and cardiovascular processes at both physiological and pathological states. This could be useful for those interested in the research field of dyslipidaemia. However, these databases are not focused on lipid metabolism (e.g., the platform CardioVINEdb for cardiovascular diseases) [97]. According to Lamaziere et al., the lack of integrated knowledge is the main current gap in the "lipidome scientific knowledge base". We need to understand how the metabolic pathways interact with each other and support as a whole integrated system the physiological mechanisms of an individual. And how changes in the regulation of these pathways can lead to the development of metabolic diseases – as perturbed states. For this, an integration of different levels of biological knowledge (e.g., transcriptome and genome) is necessary [75].

In the context of this thesis, the identification of new biomarkers able to distinguish different dyslipidaemic populations through the application of alternative approaches like, for example, ML-based models and clustering analysis, may shed further light in the metabolic pathways beyond different dyslipidaemic profiles. The exploration of genetic data related to these metabolic pathways can allow to identify additional biomarkers and genes of interest for a better discrimination between patients, thus potentially contributing for the improvement of diagnosis and treatment of dyslipidaemia. In addition, the creation of a publicly available knowledge base, for storage and visualisation of processed information, may contribute to the improvement of lipid knowledge integration.

4. Thesis aims

The main purpose of this work was to improve the distinction between monogenic and polygenic dyslipidaemia through the exploration of novel approaches in biomarker discovery, while contributing to the knowledge base of dyslipidaemia and lipid metabolism. Accordingly, four main aims were specifically defined for this doctoral thesis, as follows:

1) development of classification models, under a supervised ML approach, able to identify FH individuals with a higher specificity in comparison with the current clinical criteria, and thus improving the selection of individuals to genetic screening;
2) conduct a hierarchical clustering analysis, following an unsupervised ML approach, in order to look for potential biochemical patterns among dyslipidaemic individuals with and without FH genetic diagnosis;

3) considering the biochemical parameters that comprise the classification models trained in objective 1 and the parameters that most contribute to distinction between individuals in objective 2, identify potential biomarkers and relate them with lipid metabolism pathways in order to define a set of target genes;
4) integrate and explore gene expression patterns, molecular interactions, phenotypic and functional data associated with target genes at a new detailed knowledge base for dyslipidaemias and lipid metabolism, which should be accessible to science community.

For the first two objectives, a paediatric dataset comprised by FH+ and FH- patients was used, which is fully described in the next chapter. For the other objectives, several public databases were used to collect metabolic, genomic, transcriptomic, and functional information to establish a new knowledge base around target genes.

Chapter 2

Methods

1. Introductory note

This thesis comprises the results of work developed within three main topics, as follows: training of classification models following a supervised ML approach to improve the distinction between FH+ and FH- individuals; the application of a hierarchical clustering analysis following an unsupervised ML approach to identify biological patterns among individuals; and finally, the creation of a new knowledge base for lipid metabolism and dyslipidaemia, based on a list of target genes that was compiled taking into account the results obtained in the previous topics. For the first two work topics, classification models and hierarchical clustering analysis, a sample of 211 individuals was selected from the PFHS, which was called PFHS-ped. For the final integrative analysis, data available in public databases was used as explained forward in this chapter.

1.1. The PFHS-ped dataset: patient selection, biochemical and clinical data

The work dataset – PFHS-ped – comprises a subset of 211 children (from 2 to 17 years old) from PFHS [51] that were not undergoing statin treatment at the time of referral and for which body mass index (BMI) and a basic set of lipid parameters were available (Annex 1). PFHS was approved by the National Institute of Health Ethic Committee and National Data Protection Commission. The study protocol conforms with the ethical guidelines of the 1964 Declaration of Helsinki and its later amendments. Written informed consent was obtained from parents or legal tutors. For this study, all data were fully anonymised before analysis.

The clinical criteria to be referred to the PFHS is the SB criteria. Between 2006 and 2011, patients with LDL-C or TC levels below the cut-offs established by SB criteria were admitted to the PFHS as long as TC was above the 95th percentile for age and sex of the Portuguese population and a family history of hypercholesterolaemia was present, aiming at a better definition of the clinical criteria for FH in Portugal [48], [51]. For the purposes of this study, we decided to include these individuals in the PFHS-ped dataset to increase the number of available cases. Thus, 68% of the 211 individuals in PFHS-ped fulfil the SB clinical criteria for FH [60], while the rest present TC above the 95th percentile for their age and sex and a family history of hypercholesterolaemia [48]. All individuals were subjected to molecular study, resulting in the classification of 88 individuals as FH+ and 123 as FH-, defined respectively by presence or absence of known FH causal variants in *LDLR, APOB* or *PCSK9* genes [48]. Individuals presenting variants of unknown significance according to ACMG guidelines [52] were excluded from this study.

The PFHS-ped includes BMI, age, and an extended characterization of lipid profiles, including quantification of sdLDL, apolipoproteins (apo) A-I, A-II, B, C-II, C-III and E, and a "Lipoprint" profile measuring different subfractions of LDL-C (Table 2.1). The blood lipid profile was divided in three different levels: "Basic", "Advanced" and "Lipoprint", for commonly determined, specialized and Lipoprint test lipid parameters, respectively (Table 2.1). Biochemical characterization of "Basic" and "Advanced" lipid profiles was performed as described before [35]. Briefly, fasting blood samples were collected from individuals and TC, direct LDL-C, HDL-C, TG, apoA-I, apoB, and Lp(a) were determined for all individuals in a Cobas Integra 400 plus system (Roche) by enzymatic colorimetric and immunoturbidimetric methods. Serum levels of apoA-II, apoC-II, apoC-III, apoE, and sdLDL (sLDL-EX "SEIKEN" kit) were measured by direct quantification in a RX Daytona analyser (Randox Laboratories). The "Lipoprint" profile was obtained using the "Lipoprint LDL subfractions test" (Quantimetrix) [98]. This is a semiquantitative method that separates by polyacrylamide gel electrophoresis the different lipoprotein fractions as VLDL, IDL, LDL 1-7 subfractions (LDL subfractions 3-7 considered the sdLDL) and HDL [28], [30], [98]. For the purpose of this study, ratios that relate some lipid parameters were calculated and included as additional variables. These ratios allowed us to explore previous observations suggesting a differential contribution of TG and LDL metabolism, as well as pro-atherogenic/anti-atherogenic factors, to FH+ and FH- dyslipidaemic states (Table 2.1).

Table 2.1. Description of the biochemical parameters and ratios in each lipid profile – "Basic", "Advanced" and "Lipoprint". N/A: not applicable

Profile	Parameters		Units	Description
Basic	Biochemical	TC	mg/dl	total cholesterol
		LDL-C		low-density lipoprotein cholesterol
		HDL-C		high-density lipoprotein cholesterol
		TG		triglycerides
		Lp(a)		Lipoprotein (a)
		ApoB		Apolipoprotein B
		ApoA-I		Apolipoprotein A-I
	Ratios	ApoB/ApoA-I	N/A	pro-atherogenic vs anti-atherogenic ratio
		TG/ApoB		TG metabolism vs LDL metabolism ratio
		TC/HDL-C		pro-atherogenic vs anti-atherogenic ratio
Advanced	Biochemical	ApoA-II	mg/dl	Apolipoprotein A-II
		ApoC-II		Apolipoprotein C-II
		ApoC-III		Apolipoprotein C-III
		ApoE		Apolipoprotein E
		sdLDL.Day		Small dense LDL
	Ratios	ApoC-II/ApoC-III	N/A	anti-atherogenic vs pro-atherogenic ratio
		sdLDL/LDL-C		most atherogenic LDL in total LDL-C
Lipoprint	Biochemical	VLDL	mg/dl	Very low-density lipoprotein
		MIDA		IDL fraction A
		MIDB		IDL fraction B
		MIDC		IDL fraction C
		LDL1		Buoyant (large) LDL fraction 1
		LDL2		Buoyant (large) LDL fraction 2
		HDL.Lipo		High-density lipoprotein
		sdLDL.Lipo		Small dense LDL (fractions 3 to 7)
		IDL		Intermediate-density lipoprotein
	Ratios	VLDL/IDL	N/A	TG metabolism vs LDL metabolism ratio
		VLDL/LDL-C		TG metabolism vs LDL metabolism ratio

1.2. Categorical variables associated with the PFHS-ped dataset

BMI, age, and the biochemical parameters present in Table 2.1. are all quantitative variables that describe PFHS-ped and were considered for both modelling and clustering analysis. PFHS-ped also contains a set of categorical variables that were only used as supplementary variables in the clustering analysis (see section 3). These variables are the following: "Class", classification based on FH genotype (FH+ or FH-); "Gender" (female or male); "Activity class", according to the percentage of molecular activity that is kept by the affected gene allele; "Gene", the affected gene in FH+ individuals (*LDLR, APOB, PCSK9*); "SB criteria", concerning the fulfilment of TC and/or LDL-C cut-offs from Simon Broome clinical criteria (yes or no); "Lipoprint profile", according to low or high concentration of sdLDL in serum

measured by Lipoprint assay (profile A or B, respectively); and "LDL-C score", a polygenic risk score associated with LDL-C levels and based on a panel of six SNPs, as performed by Mariano et al. [69]. The categories of "LDL-C score", "Activity class" and a new variable (called "BMI class") were established for the purpose of this study. The variable "LDL-C score" comprises four categories that correspond, from a high to a low polygenic contribution, to the following: ≥0.9, 0.9-0.7, 0.7-0.5, <0.5. Regarding "Activity class", variants were divided in the following categories: null variants, presenting less than 2% of activity compared to wild type allele; defective variants with different degrees of molecular activity corresponding to three categories (2-20%, 20-40%, 40-65%); null_pred, variants predicted to be null according to *in silico* analysis; def_pred, variants predicted to be defective according to *in silico* analysis. Concerning the new variable, "BMI class" comprises the classification of BMI according to gender and age, following the percentiles for children and adolescents from the World Health Organization (WHO) [99], [100]. Accordingly, "BMI class" includes the following categories: severe thinness, thinness, normal, overweight, obesity.

2. Training of classification models that improve distinction between FH+ and FH- individuals

All the analysis was developed using R software (version 3.4.3) [101]. The *caret* package for ML [102] was used to train classification models based on logistic regression, and a resampling scheme of three times cross validation was applied to estimate model accuracy. Accordingly, data was randomly divided in two sets of 60% and 40% of the subjects defining the training and the testing sets, respectively. The training set was used for model generation and the testing set was used for posterior validation. The Bayesian generalised linear model (bayesglm) was applied on the training set using the *train* function of *caret* package [103]. To avoid overfitting the number of parameters considered for model training was reduced using three different methods available in caret : 1) exclusion of one element of a pair of highly correlated parameters (cut-off = 0.8; "cor models"); 2) ranking of parameters by importance based on a ROC analysis of each parameter with only the top 3 variables selected for model training ("Imp models"); 3) recursive feature elimination (RFE) of parameters ("RFE models"); when more than five parameters were selected by RFE, models were trained both with the top 5 and with all parameters. Additional "RFE models" were generated after removal of highly correlated parameters. To identify highly correlated parameters, *cor* function was applied using Kendal's tau method for mixed and tied data [104]. Following model training, their predictive performance was assessed on the testing set. The *predict* function of *caret* was used along with

confusionMatrix to acquire the main statistics regarding model performance, which were organised in a table using *broom* package [105]. The *pROC* package [106] was used to measure the area under the ROC curve (AUC) regarding individuals classification.

Model ranking criteria were defined as follows: 1) sorting by AUC (highest to lowest) reflects a better performance regarding the relation between specificity and sensitivity; 2) k values above 0.4 correspond to a moderate agreement between observed and predicted classes [107]; 3) reduced number of parameters (≤5) for model simplicity; and 4) highest sensitivity (≥0.7) cut-off values. Criteria were applied in that order to generate a ranked list of the best models, 11 in total. The Akaike information criterion (AIC) was used to eliminate one model to restrict the final list to a "top 10".

3. Identification of different dyslipidaemic profiles among individuals by a hierarchical clustering analysis

All data analysis was performed using R software (R version 3.4.3) [101]. For the hierarchical clustering of principal components (HCPC) analysis, we used *PCA* and *HCPC* functions of the package *FactomineR* (version 1.41) [108].

The *PCA* function performs PCA with the possibility of adding supplementary individuals and variables, both quantitative and categorical. This supplementary data does not contribute to PCA itself but can be useful for results interpretation [109]. In the current data analysis, "Class" was used as a discriminant factor while observing the distribution of individuals classified as FH+ or FH- across the clusters. For the "All" subset, since it was the subset that showed the most well-defined cluster partition (see results), other categorical variables besides "Class" were included as supplementary data to test their association with clusters, such as "BMI class", "Activity class", "Gender", "Gene", "SB criteria", "Lipoprint profile", and "LDL-C score".

The *HCPC* function performs an agglomerative hierarchical clustering on the results of PCA, by using the PCA coordinates of individuals to measure the distance between them. For this, HCPC uses Ward's minimum variance method that aggregates clusters when it translates in a minimum of within-variance growth, thus contributing for homogenous clusters [110]. In this study, the hierarchical clustering was performed using the first five dimensions of PCA (default option), which explain most of the data variance. The obtained hierarchical tree, also known as dendrogram, was cut at the suggested level, corresponding to a partition in three clusters. HCPC output also includes a description of the clusters by individuals, variables, and dimensions; assignment of a cluster for each patient; other graphic visualisations besides the dendrogram, such as cluster maps [111]. The functions *fviz_dend* and *fviz_cluster* of the package *factoextra*

were applied to improve the visualisation and interpretation of dendrograms and cluster maps [112]. Both functions allow an enhanced ggplot2-based visualisation of the plots [113].

For the characterization of clusters by quantitative variables, the correlation ratio was directly measured by the HCPC algorithm between each variable and the cluster partition (i.e., cluster assignment of individuals, called "cluster variable" by HCPC), which was followed by a student's t-test that determined the variables whose correlation ratio was significantly different from zero, and thus had a significant contribution for cluster partition. In addition, HCPC measured the average of a variable in the cluster (named as "mean in category") and in the whole subset (named as "overall mean"), including the associated standard deviations. Then, the HCPC algorithm applied a student's t-test to enquire, for each variable, if the difference between the mean in category and the overall mean was statistically significant under the confidence level of 95%. A positive or negative value of the test statistics indicates if the mean in category is greater or lower than the overall mean, respectively. Regarding categorical variables, a chi-squared test was performed to assess the association between each variable and the cluster partition. Then, for each variable category, a statistical test was performed to check if the proportion of individuals within the category that belong to a given cluster was significantly different from the proportion of individuals assigned to the same cluster that belong to this category. This allowed us to assess which categories were underrepresented or overrepresented in each cluster. For a description of clusters by individuals, the distance between each individual coordinate and the gravity centre of the assigned cluster was measured to assess the paragons, which means the individuals that most characterise each cluster because they are the closest to the cluster centre. In addition, measuring the distance between each individual coordinate and the gravity centre of other clusters allowed the assessment of the most specific individuals of each cluster, which correspond to the farthest individuals to the centre of the other clusters. For cluster characterization by dimensions an identical analysis to that carried out for quantitative variables was performed, considering each PCA dimension a quantitative variable composed by individual coordinates. The methodology behind HCPC analysis is explained in more detail by Husson et al. [110] and Lê et al. [108].

For prediction of class probabilities of individuals from the "All" subset, the best trained model of the top10 models (Imp_B model) was applied in that subset. The basic function *predict* was used with the argument "type" set for probabilities. Then, each of the 78 individuals was assigned to a predicted class (FH+ or FH-) and to the probability of belonging to each of the two classes. The sum of both probability values was equal to 1. In addition, the difference between probabilities (called Δprob) was measured for each individual, using absolute values

of probability. Categories were established for Δprob taking into account that the lower the value, the more ambiguous is the classification of individuals.

For all the figures showing dendrograms, *dplyr* [114] and *ggplot2* [115] packages were used to assign a different colour to each variable category and plot the variables aligned with individuals by their order of appearance in the dendrogram.

4. Creation of a new lipid knowledge base directed to dyslipidaemia

4.1. Building the "MylipidgenesKB" knowledge base

For this section of integrative data analysis publicly available databases were used, firstly to establish a list of genes associated to dyslipidaemia and lipid metabolism, and secondly to add valuable information regarding each of these target genes. The collected information included gene expression data, associated GWAS traits, and functional characterization given by gene ontology (GO) terms. For the establishment of the target gene list, a set of keywords - based on literature review and previous results of both modelling and clustering analysis, was used to search and select a list of metabolic pathways of interest on Wikipathways [116]. All the genes in each of the selected pathways were joined in a single list of target genes. The gene IDs were standardised in order to have all the genes from this list with the same ID system (Ensembl). For this, the DAVID online resource "Gene ID Conversion Tool" was used [117], [118], which also allowed to add and/or check the full name and symbol of each gene. In addition, pathway assignment was saved for these genes, with each of them being associated to one or more of the previously selected metabolic pathways.

Expression data was taken from Genotype-Tissue Expression (GTEx) database [119], namely GTEx v7 dataset that comprises median transcript per million (TPM) measures of RNA-seq expression data of 56202 genes for 51 tissues and two modified cellular lines, which are commonly used in scientific research. This dataset involved 11688 RNA-seq samples from 714 individuals. For the purposes of this study, the two cellular lines – Epstein-Barr virus transformed cells and transformed fibroblasts, both produced *in vitro* – were not considered as tissues for the estimation of median gene expression in the human transcriptome. This transcriptome estimation, called in the text ahead simply as "Transcriptome", was acquired by dividing the mean for the sum of expression values of each gene across 51 tissues. For tissues of interest, namely liver and small intestine, cut-offs and respective expression categories were determined as follows: set 0.1 TPM as the minimum value of gene expression below which a gene is considered to be in the null expression category, apply a logarithmic transformation

(base 10) of data and check its distribution by plotting a histogram, establish cut-offs and categories taking into account data distribution as well as median and mean values.

For the collection of GWAS information, a set of keywords associated with lipid metabolism and dyslipidaemia (Table 2.2) was selected based on the literature review. These keywords were searched on the NHGRI-EBI GWAS catalog [120] to find out if those were considered as phenotypic traits within any GWAS study, which was followed by a search for the genes associated with each trait/keyword. For this, the functions *get_traits* and *get_associations* of the R package *gwasrapidd* [121] were used, respectively. Afterwards, the list of genes associated with each of the nine identified GWAS traits was compared with the list of target genes, thus allowing to find the traits associated with each target gene.

Table 2.2. Keywords used in the trait search on the NHGRI-EBI GWAS catalog considering metabolites, pathways, and conditions of interest.

Keyword	**Class**
Hypercholesterolaemia(s)/hypercholesterolemia(s)	Disease
Familial hypercholesterol(a)emia	Disease
Lipid disorder(s)/disease(s) metabolism	Disease/pathway
Lipoprotein disorder(s)/disease(s)	Disease
High cholesterol/cholesterol (measurement)	Metabolite
HDL(-C)/LDL(-C) measurement/metabolism	Metabolite/pathway
Triglycerides measurement/metabolism	Metabolite/pathway
(Premature) cardiovascular disease(s)/disorder(s)/risk	Disease
Atherosclerosis	Disease
Lipoprotein measurement	Metabolite
Coronary artery disease	Disease
Hypertriglycerid(a)emia	Disease

To show the representation of each GWAS trait among target genes, a bar plot was drawn considering the number of target genes associated with each trait, using the *plotly* package [122]. The sum of the number of genes represented in the plot is not equal to the total number of target genes, since not all target genes presented associations with at least one of the nine identified traits, and there were target genes associated with more than one trait.

The whole set of GO terms associated to target genes were collected for GO domains "molecular function" (MF) and "biological process" (BP), using *getBM* function of package *biomaRt* [123], [124] to remotely access Ensembl database for *Homo sapiens* dataset. The number of target genes associated with each GO term was measured to help in the selection of the most representative terms. Two sets of GO terms, lipid-specific and others, were manually selected within each GO domain from the full lists of GO terms obtained with *biomaRt*. The

selection of lipid-specific GO terms was based on previous knowledge regarding lipid metabolism and dyslipidaemia, thus aiming to select the most representative GO terms of target genes considering the biological context of this thesis. From the list of the remainder GO terms, not considered to be lipid-specific, the most representative GO terms were selected based on the number of associated target genes and the hierarchical relations between terms, acquiring a list of terms that was called "Other GO terms". Both lists of terms (lipid-specific and other GO terms) comprised "parent" and "child" terms, and thus the terms were grouped according to their hierarchical relations in the GO graph, where "parent" terms refer to the nodes closer to the graph's root and "child" terms refer to those closer to the leaf nodes.

Considering the new knowledge base, this is organised in three sections (Figure 2.1). Firstly, one main table comprising all target genes and associated information (i.e., metabolic pathways, gene expression data, associated GWAS traits and GO terms), with signalization of those that are core genes. The second section is composed by a set of auxiliary files of gene associated information, including the lists of lipid-related GO terms and other GO terms organised by their hierarchical relations for each GO domain, and the list of target genes with the number of GWAS traits by gene. The third section comprises the gene interaction networks that were built (see section 4.2) taking into account the genes associated to the selected GWAS traits, and the core genes. These networks have associated tables with nodes and edges attributes that allow reconstructing each network. The entire knowledge base occupies a total size of 3,22 MB.

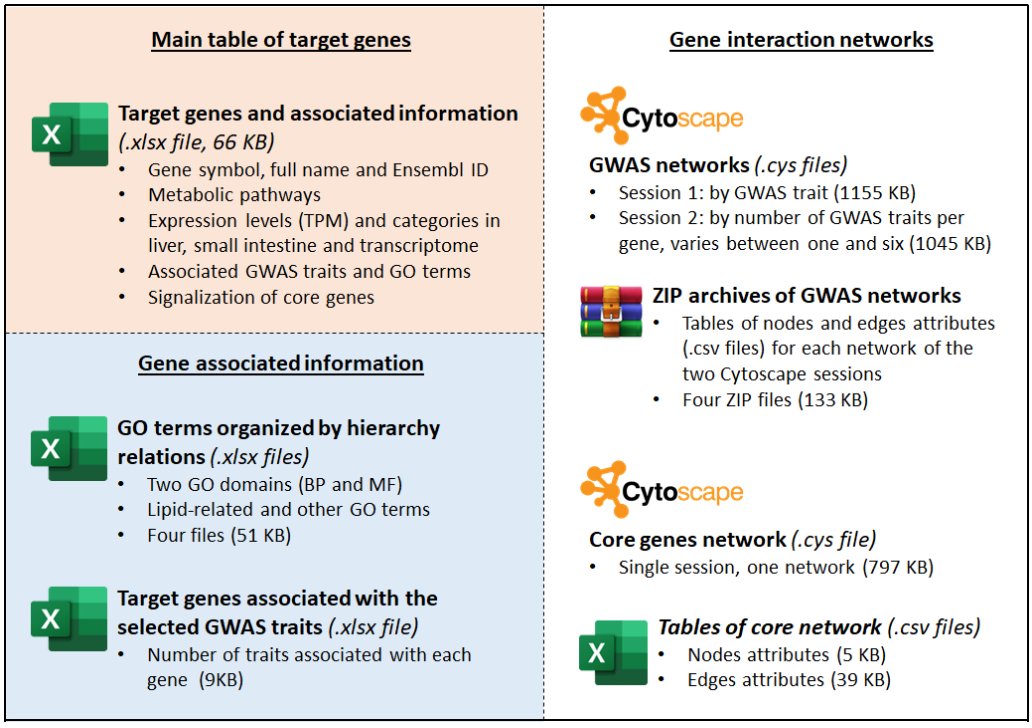


Figure 2.1. Organisation of the new knowledge base in three sections and the files that composed each section, including their size and associated data.

4.2. Development of a shiny app to host the new lipid knowledge base

The "MylipidgenesKB" website was built in R (version 4.0.3) [125] using *shiny* package [126], with the definition of a ui (i.e., user interface) and a server that are set up together by the *shinyApp* function. The website follows a dashboard structure, which comprises a header, sidebar, and body. The header contains the title and logo of the website, the sidebar presents a menu that allows the user to select among the different sections of information, the body comprises the main contents of the website. For this, functions *dashboardSidebar* and *dashboardBody* from *shinydashboard* package [127] were applied. The functions *dashboardPagePlus* and *dashboardHeaderPlus* of the *shinydashboardPlus* package [128] allowed the option to partially collapse the sidebar menu showing only the respective icons. From package *shinyWidgets* [129], function *searchInput* and *multiInput* were used for the search bar of the homepage and for the gene selector field of the "tissue expression" section, respectively. In addition, the *selectInput* function of *shiny* package [126] was used for the tissue selector field in this last section. For the output table of homepage, gene list and GO terms tables, package *DT* [130] was used with *datatable*, *dataTableOutput* and *renderDataTable* functions. Regarding GO terms tables, the extension "Buttons" was used as argument of *datatable* function (offering buttons "copy", "csv", "print"), and "dom" argument was set as "Bft" (i.e., buttons, filtering input, table). For output table of homepage and gene list table, the "dom" argument was set as "t" (i.e., only the table) and "lftp" (i.e., length changing input control, filtering input, table, pagination control), respectively. For "gene list" and selected data in "tissue expression" sections, *downloadButton* and *downloadHandler* functions of *shiny* package [126] were used to set a download button for the respective data (in .csv format). To achieve the interactive heatmap for gene expression patterns, functions *plotlyOutput* and *renderPlotly* from *plotly* package [122] were used taking into account the input data for tissues and genes. For text of "About" and "GWAS traits" sections, the function *includeHTML* of *shiny* package [126] was used to include .txt files containing the text formatted in HTML language.

Regarding the "GWAS traits" section, each gene interactions network was created in Cytoscape (version 3.8.2) [131], using the GeneMANIA app (version 3.5.2) [132] for local search of interactions among the previously identified GWAS genes (i.e., target genes with associated GWAS traits). In this GeneMANIA search, four types of interactions were selected to establish the networks, including co-expression, pathway, physical and genetic interactions. This process was split in two different sessions of Cytoscape as follows: in session "by trait", a network was established for each of the nine previously identified traits; in session "by trait number", a

network was created for genes sharing the same number of associated traits (e.g., two traits network only comprises GWAS genes with two associated traits), and an additional network was established to include all GWAS genes independently of the number of associated traits. Then, each network session was exported from Cytoscape as a full web application (interactive Cytoscape.js webpage) in a .zip file, containing a HTML page with a network interactive viewer that allows users to select a network and a layout. This HTML page was included as an iframe object in a new HTML page, prepared in a R HTML file, together with network legends and links for download of edges and nodes tables. This new HTML page was included in the shiny app using *includedHTML* function from *shiny* package [126]. To split these two groups of networks in two different subsections within the "GWAS traits" section, the function *tabBox* from *shiny* package [126] was applied with two panels. The same approach, from Cytoscape to shiny app, was used to establish the core genes network. The designation "core" was only used in the shiny app "MylipidgenesKB" for candidate genes.

Chapter 3

Results

1. Training of classification models that improve distinction between FH+ and FH- individuals

The results of this section are published in the following article: Correia, M., Kagenaar, E., van Schalkwijk, D.B. *et al.* Machine learning modelling of blood lipid biomarkers in familial hypercholesterolaemia versus polygenic/environmental dyslipidaemia. *Sci Rep* **11,** 3801 (2021). https://doi.org/10.1038/s41598-021-83392-w

In this work we used a ML approach to explore the paediatric subset of the PFHS 2018 dataset update (PFHS-ped) to develop novel models that can integrate data from multiple biomarkers and achieve a reliable discrimination between individuals. Our systematic exploration of available lipid parameters resulted in the development of several models that can robustly classify subjects into FH+ or FH− classes. Some of the models have parameters not routinely used in clinical practice but that are commercially available. Notwithstanding, models comprising only the standard lipid parameters used in the clinic also achieved a relatively good performance. Our results provide an approach for improving the yield of genetic screening programs while showing distinct biochemical backgrounds in monogenic and polygenic hypercholesterolaemia.

1.1. Definition of PFHS-ped data subsets for exploratory modelling of extended lipid profiles

Given that the available information on lipid parameters varied between individuals and considering the three lipid profiles defined for this study - “Basic”, “Advanced”, and “Lipoprint”, we began by establishing distinct data subsets regarding all the possible combinations of these profiles (Figure 3.1). A detailed description of the seven data subsets is available in the Annex 2.

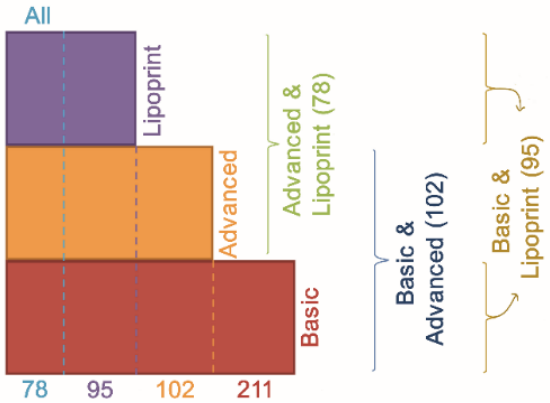


Figure 3.1. Data subsets used for model training. Figure shows how PFHS-ped was divided into smaller subsets, identified by a colour-coded size (number of individuals) and name, according to the available biochemical parameters for each individual.

As depicted in Figure 3.1, the number of individuals across subsets varies between 78 and 211. Although relatively small, these numbers have been previously used in conjugation with ML approaches to derive valuable insights into complex biological problems [133]–[136]. We therefore set out to systematically search for the best model to discriminate between FH+ and FH- individuals using these different combinations of lipid parameters. The different number of individuals between subsets creates a challenge regarding results comparison. Thus, the models generated for each subset were further tested using a minimal dataset composed of 78 subjects, as explained below.

1.2. Systematic training of models to distinguish FH+ and FH- subjects using extended lipid profiles

We began by training models using all available parameters in each subset. These "pilot models" provided a rough overview of the behaviour of the different parameters in our data subsets but presented a very low performance as assessed by their sensitivity and specificity values (Annex 3). This suggested an overfitting problem, likely linked to the high number of variables present in the dataset versus the total number of observations. To overcome this issue, we applied three different commonly used methods to reduce the number of parameters considered for model training (see methods). This systematic approach resulted in a total of 35 models belonging to one of three categories: "cor models", "Imp models", and "RFE models" (Figure 3.2).

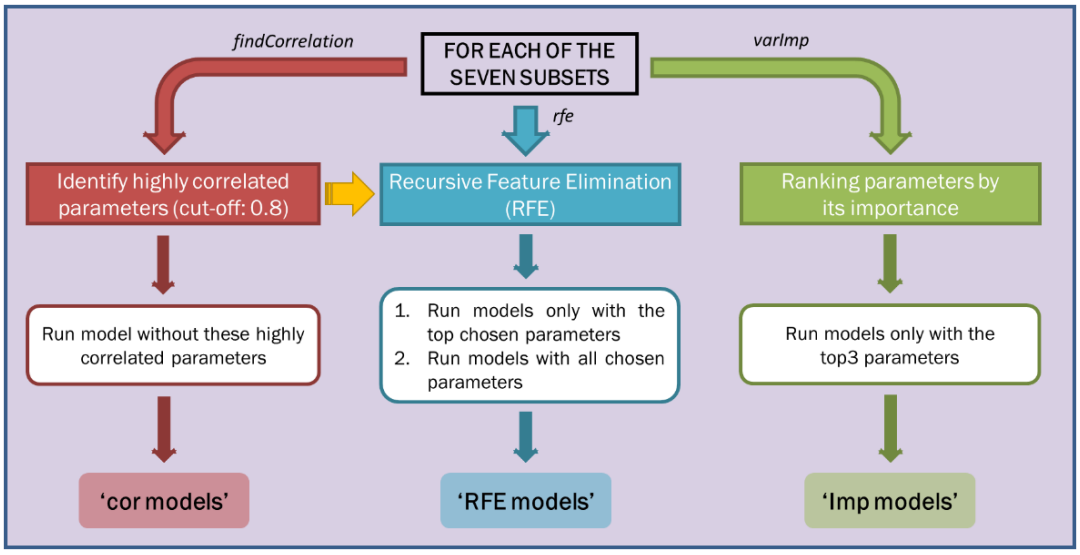


Figure 3.2. Modelling workflow using three methods to avoid overfitting, producing three groups of models: "cor models", "RFE models", and "Imp models". This approach was applied individually for each subset.

Interestingly, a trend towards the selection of parameters from the "Advanced" and "Lipoprint" profiles as the most relevant for distinguishing FH+ from FH- subjects (Annex 3) was observed. Considering the relatively small size of the corresponding data subsets, we decided to investigate whether it could be influencing the perceived contribution of "Advanced" and "Lipoprint" parameters in our models.

For this purpose, we repeated our analysis (Figure 3.2) using the biochemical parameters available for each data subset restricting the number of individuals to 78. This number corresponds to the smaller sized subset used in this study (the "All" subset), which comprises the subjects that present measures for all biochemical parameters. Two different approaches were followed: train all the models with the same 78 subjects from the "All" subset; or use a random selection of 78 subjects from the corresponding data subset. This analysis confirmed that parameters from the "Advanced" and "Lipoprint" profiles contribute to a better discrimination between FH+ and FH- status independently of the training set (Annex 3).

Through careful inspection of all models regarding variable importance and correlation, we noticed that a group of four parameters (LDL1, apoC-III, TC/HDL-C and sdLDL.Day) consistently appeared as highly relevant for the discrimination between FH+ and FH- individuals. However, none of the trained models used this small group of parameters as the only predictors. Such models could be relevant for clinical purposes given their comparative simplicity. Therefore, we decided to train two additional models including only these selected parameters (Sel1 and Sel2, Table 3.1). Given that BMI and age are likely to influence the lipid profile of subjects [35], [137], we further conjugated these parameters with them (models Sel3 and Sel4, Table 3.1). Given the fact that these "selected models" comprise parameters from different lipid profiles, they were trained on the "All" subset.

Table 3.1. Identification of the manually selected parameters that comprise each of the four "selected models". N: number of individuals; Np: number of parameters.

Model	N	Np	Parameters
Sel1	78	3	LDL1 + ApoC-III + TC/HDL-C
Sel2	78	4	LDL1 + ApoC-III + TC/HDL-C + sdLDL.Day
Sel3	78	5	LDL1 + ApoC-III + TC/HDL-C + BMI + Age
Sel4	78	6	LDL1 + ApoC-III + TC/HDL-C + sdLDL.Day + BMI + Age

Altogether, a total of 67 models were generated during this analysis (Annex 3). Given that the presence of models with highly correlated parameters does not contribute substantially to new insights into the biological background of dyslipidaemia, we identified all models containing

any pair of parameters whose correlation was equal to or higher than |0.6|. For this purpose, we generated a correlation plot for all parameters used during modelling analysis (Figure 3.3).

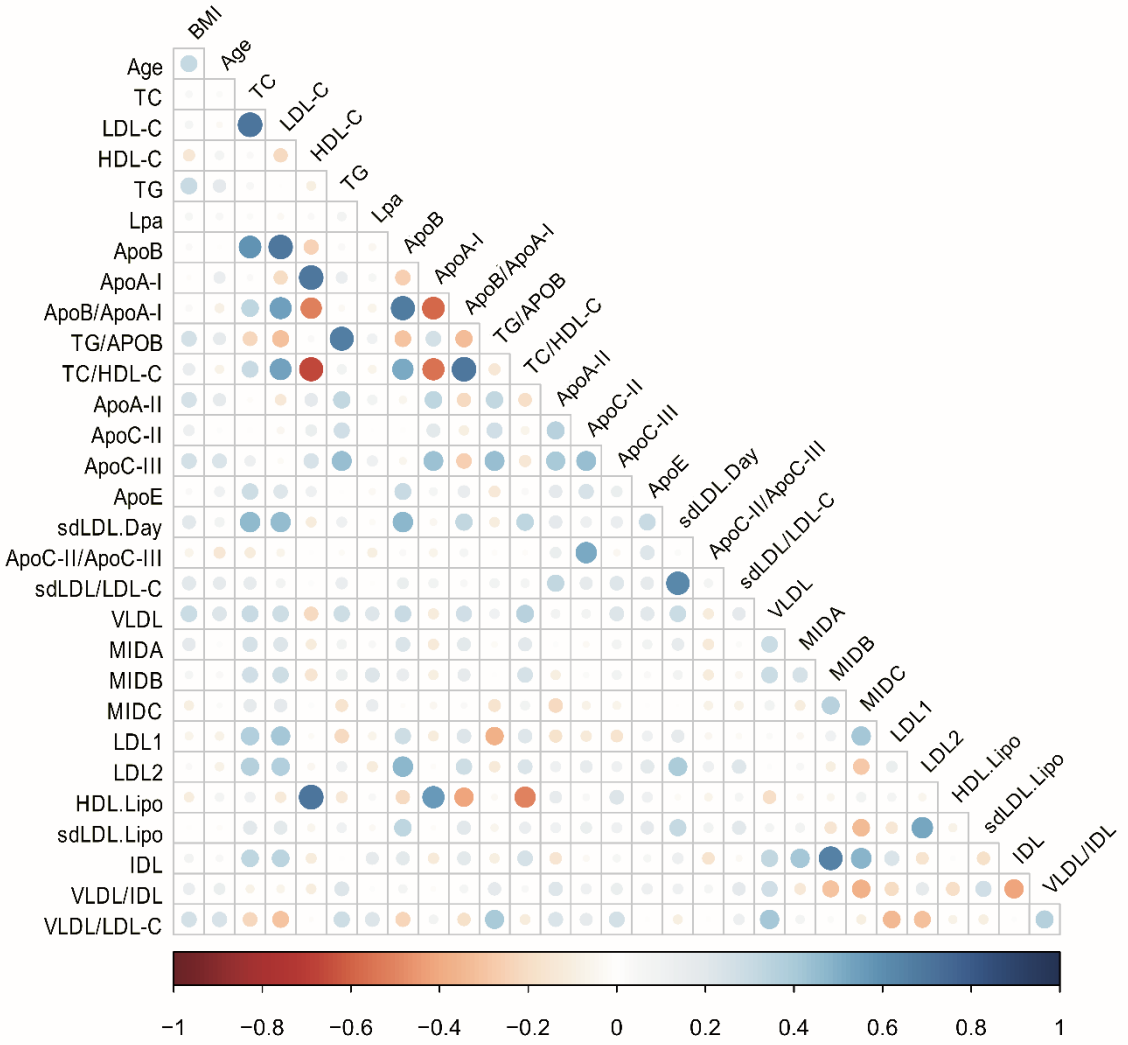


Figure 3.3. Correlation plot for the dataset parameters. Negative and positive correlations are presented in red and blue, with darker colours corresponding to higher absolute values, according to the scale.

A total of 14 pairs of highly correlated parameters were identified, 12 of which belong to the "Basic" profile. These pairs were found in 32 out of 67 trained models and were thus discarded from further analysis.

1.3. Extended lipid profiles contribute to distinguish FH+ and FH- subjects

Following model training, testing datasets were used to assess model performance and corresponding descriptive statistics were determined. We established a set of ranking criteria to apply to the 35 final models, with cut-off values defined considering the properties and

observed range for each statistic (see methods). We used this approach to retain only the top 10 models (Table 3.2).

Table 3.2. Top ranking models and performance. N: number of individuals; Np: number of parameters; Acc: accuracy; k: Cohen's kappa coefficient; Sens: sensitivity; Spec: specificity; TP: number of true positives; FN: number of false negatives; FP: number of false positives; TN: number of true negatives; AUC: area under the ROC curve.

Model	Subset	N	Np	Parameters	Acc	k	Sens	Spec	TP	FN	FP	TN	AUC
Imp_B	Basic	211	3	LDL-C + ApoB/ApoA-I + TG/ApoB	0.84	0.67	0.91	0.86	32	3	7	42	0.92
RFEct_BL	Basic & Lipoprint	95	4	TG/ApoB + TC/HDL-C + TC + LDL1	0.84	0.64	0.83	0.92	10	2	2	23	0.91
Sel3	All	78	5	LDL1 + ApoC-III + TC/HDL-C + BMI + Age	0.77	0.49	0.82	0.90	9	2	2	18	0.89
RFEct_A	All	78	5	LDL1 + TC + ApoA-II + MIDC + TC/HDL-C	0.77	0.46	0.82	0.80	9	2	4	16	0.88
RFE78ct_BL	Basic & Lipoprint	78	5	TC + TC/HDL-C + MIDB + MIDC + LDL1	0.74	0.41	0.82	0.85	9	2	3	17	0.88
RFE78t_B	Basic	78	2	LDL-C + ApoB/ApoA-I	0.81	0.59	0.82	0.85	9	2	3	17	0.87
Sel1	All	78	3	LDL1 + ApoC-III + TC/HDL-C	0.77	0.47	0.82	0.90	9	2	2	18	0.87
Imp_AdL	Advanced & Lipoprint	78	3	ApoA-II + ApoC-III + LDL1	0.77	0.47	0.73	0.75	8	3	5	15	0.76
RFE78t_Ad	Advanced	78	5	ApoA-II + ApoC-II + ApoC-III + sdLDL.Day + BMI	0.77	0.49	0.91	0.60	10	1	8	12	0.75
RFE78ct_Ad	Advanced	78	5	Age + ApoA-II + ApoC-II + ApoC-III + sdLDL.Day	0.85	0.66	0.73	0.65	8	3	7	13	0.75

The two best ranked models were the Imp_B and RFEct_BL models, trained with the "Basic" and the "Basic & Lipoprint" subsets, respectively. Among the top 10, these models presented the highest AUC values combined with the best k metrics (Table 3.2), revealing a substantial agreement between observed and predicted classification of subjects [107]. These models further display the best association between sensitivity and specificity, with Imp_B performing better for sensitivity and RFEct_BL for specificity. Of note, eight of the top 10 models were trained using at least one parameter of the "Advanced" and/or "Lipoprint" profiles. The Lipoprint measurement for LDL1 is present in six of these models. The other models

(RFE78t_Ad and RFE78ct_Ad) include sdLDL.Day, apoA-II, apoC-II and apoC-III values from the "Advanced" profile. The models that were trained using only parameters from the "Basic" profile include the apoB/apoA-I ratio in addition to LDL-C (Imp_B and RFE78t_B). The Imp_B model further includes the TG/apoB ratio. Of note, out of the 32 models removed due to the presence of highly correlated variables, only 5 were above the top 10 cut-off criteria and contained the same variables selected in top 10 models, supporting our choice to discard them (not shown).

In summary, the comparative analysis of model performance revealed that the integration of lipid parameters from different profiles following a ML-based approach can support a robust discrimination between FH+ and FH- subjects (Table 3.2). Moreover, our results suggest that biochemical parameters not commonly used in clinical practice, but available commercially, may provide important information towards this distinction, namely contributing to a higher specificity.

1.4. Modelling of TC and LDL-C levels improves identification of FH+ individuals in comparison to clinical cut-offs

The biochemical parameters and cut-offs of the SB criteria are widely used to identify candidate FH individuals and refer them for therapy and genetic testing [35]. Of note, only about 60% of the PFHS-ped individuals that fulfilled these criteria were actually FH+, whereas three FH+ individuals were found among the 67 that had TC or LDL-C values below these cut-offs.

Given that the SB criteria are based on two simple biochemical parameters, we decided to train two models exclusively using TC and LDL-C and assess their ability to correctly distinguish between FH+ and FH- individuals ("SB models", Table 3.3).

Table 3.3. Performance of models trained with SB criteria parameters. Column names as defined in Table 3.2 legend.

Model	Subset	N	Np	Parameters	Acc	k	Sens	Spec	TP	FN	FP	TN	AUC
SB_B	Basic	211	2	TC + LDL-C	0.80	0.57	0.77	0.82	27	8	9	40	0.89
SB_BL	Basic & Lipoprint	95	2	TC + LDL-C	0.81	0.56	0.67	0.84	8	4	4	21	0.84

These models were trained using all the PFHS-ped subjects or just the "Basic & Lipoprint" subset used to train the second best ranked model (Table 3.2). The resulting SB models had a weaker performance when compared to top 10 models trained on the same subsets (cf. Table 3.2 and Table 3.3). To explore the differences between SB models and the two best ranked

models, we used them to classify 50 individuals randomly selected from the "Basic & Lipoprint" subset (Table 3.4).

Table 3.4. Comparison of classification performance between the best two ranked models, "SB models" and SB criteria for the same universe of 50 individuals (randomly selected from "Basic & Lipoprint" subset). PPV: positive predictive value; NPV: negative predictive value.

		Specificity	Sensitivity	PPV	NPV
Models	Imp_B	0.89	0.93	0.78	0.97
	RFEct_BL	0.97	0.87	0.93	0.94
	SB_BL	0.83	0.73	0.65	0.88
	SB_B	0.86	0.80	0.71	0.91
SB criteria		0.49	1	0.45	1

Specificity, sensitivity, and the positive and negative predictive values (PPV and NPV, respectively) were calculated for the predictions made by these models, as well as for the FH+/FH- classification according to SB criteria cut-offs (Table 3.4). As expected, SB criteria have a very high sensitivity and NPV. However, they are extremely unspecific, with a high likelihood of selection of FH- patients for genetic testing. SB models can considerably improve on this, although they present a lower sensitivity in comparison to SB cut-offs. However, in contrast with SB cut-offs, these models present a very good balance between sensitivity and specificity (Table 3.4). The two top-ranked models trained with the extended lipid profile can achieve very good PPVs while keeping acceptable values for sensitivity and NPV.

These results emphasise how modelling approaches can improve patient classification compared to the use of strict cut-off values. The reduced performance of SB models in comparison to top 10 models supports our suggestion that extended lipid parameters contain relevant biological information for an improved classification of FH+ and FH- individuals.

1.5. Implementing the best-ranking models in a clinical setting

Our top 10 models can be easily used in clinical practice to prioritise patients for genetic testing. Clinicians can access the different models and select the one that better suits their practice, in the following link: https://github.com/GamaPintoLab/FH-Models-.git. Models can be grouped into three different categories, depending on the availability of parameters required to run them. A first set of models, including the best ranked model, require biochemical parameters that can be provided by most clinical laboratories. Other models include additional values for apoA-II,

apoC-II, apoC-III, sdLDL.Day, which are only available in more specialised clinical laboratories, while the final set of models relies on "Lipoprint" parameters LDL1, MIDC or MIDB, a method that is currently for research use only. We provide an Excel file for simple implementation of the two best ranked models (Table 3.2) and the SB_B model, which classifies patients as FH+ or FH- upon introduction of the required parameter values. In addition, all top 10 models can be downloaded and applied to a new dataset using R software.

1.6. Biochemical parameters identified through machine learning provide novel insights into the biology of hypercholesterolaemia

Given the high efficiency of the trained models in discriminating FH+ and FH- individuals, we next addressed if the parameters selected in the best two models (Table 3.2) could provide novel insights into metabolic differences between these groups. Figure 3.4 shows the mean and distribution of these parameters and Figure 3.5 summarises their connection to lipoprotein metabolism.

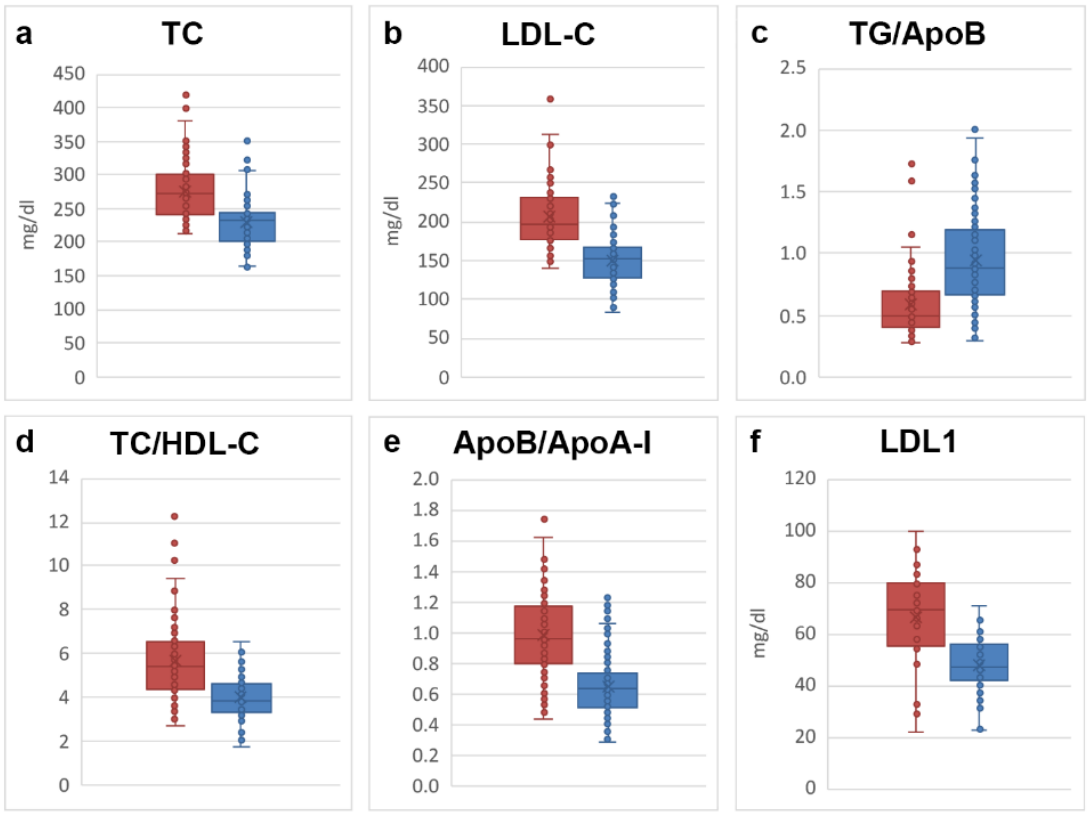


Figure 3.4. Box and whiskers plots with distribution of individual values for the parameters used by the two top ranking models according to patient classification as FH+ (red) or FH- (blue). a) total cholesterol (TC); b) LDL-cholesterol (LDL-C); c) ratio between triglycerides and apolipoprotein B (TG/ApoB); d) ratio between total cholesterol and HDL-cholesterol (TC/HDL-C); e) ratio between apolipoprotein B and apolipoprotein A-I (ApoB/ApoA-I); f) buoyant (large) LDL fraction 1 (LDL1). All individuals with measured TC, LDL-C, and LDL1 were used for the plots a), b) and f), respectively. For ratios, all individuals with measurements for both parameters were used.

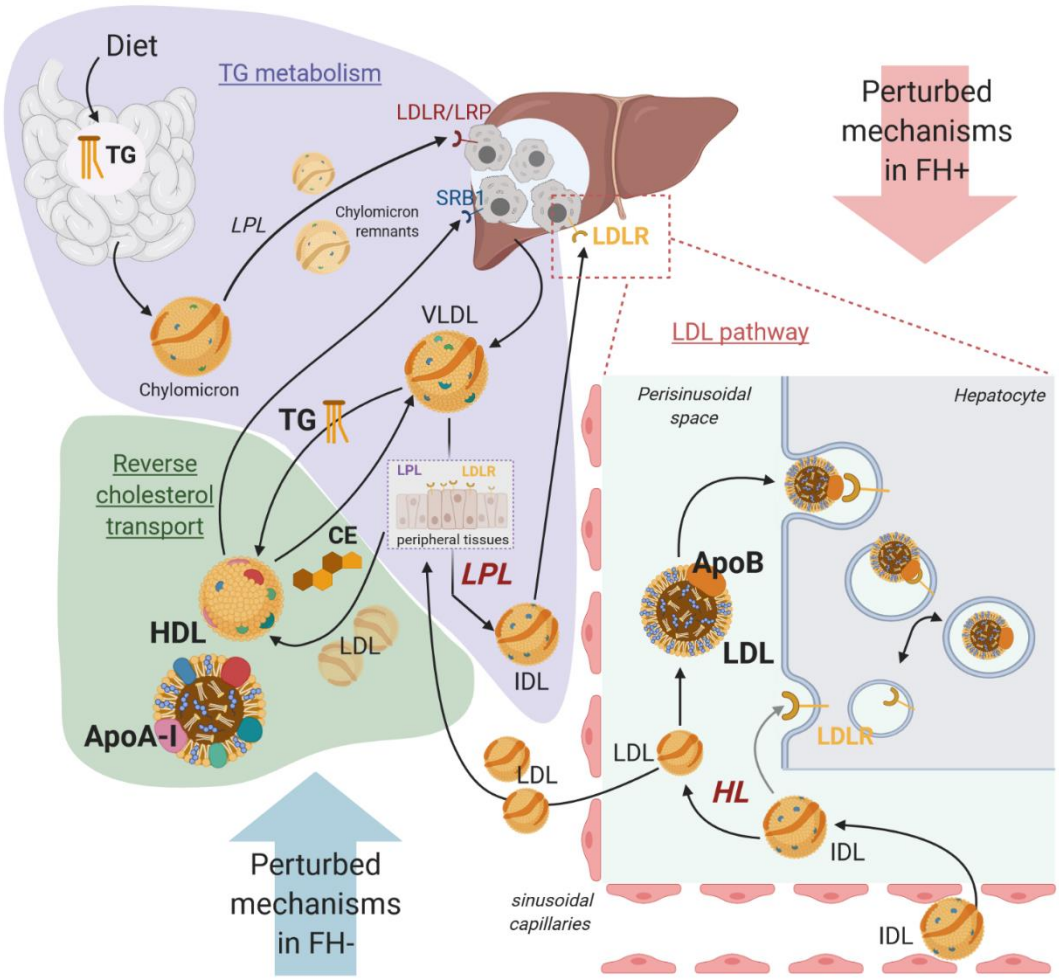


Figure 3.5. Main pathways involved in lipoprotein metabolism. Inferred differences between FH+ and FH- individuals are indicated by arrows. Created with BioRender.com.

We find that TC and LDL-C tend to have higher values for FH+ compared to FH- subjects (Figure 3.4a and 3.4b). This is expected because FH+ subjects present single-gene variants that disrupt the clearance of LDL particles by the liver [138], leading to the build-up of LDL-C in circulation, and thus of TC. However, given the high variability of these parameters and their largely overlapping distribution, they are not enough to properly discriminate FH+ cases.

We further find that the TG/apoB ratio is lower for FH+ compared to FH- subjects (Figure 3.4c). In this regard, it is interesting to consider that hypercholesterolaemia in FH- subjects is likely to have environmental influence, such as cholesterol and TG-rich diets [111]. Thus, given both the lower clearance of LDL-associated apoB in FH+ patients and higher blood TG levels in FH- subjects (Annex 2), this ratio provides additional discriminating power. Of note, LDL-C and apoB levels are highly correlated, as expected (Figure 3.3).

The TC/HDL-C and apoB/apoA-I ratios are higher in FH+ compared to FH- subjects (Figure 3.4d and 3.4e). Higher TG availability in FH- subjects should lead to a production of more and

"bigger" VLDL particles [139], which results in their increased lipolysis through LPL (Figure 3.5). The released cholesterol is transported back to the liver as HDL, raising HDL-C and apoA-I concentrations in FH- individuals. This mechanism is plausible because LPL gain-of-function and loss-of-function polymorphisms lead to higher and lower HDL-C, respectively [140]. As TC and apoB levels follow an opposite trend, being increased in FH+ individuals (see discussion above), these ratios again afford higher discrimination than the individual parameters.

We consistently find a higher LDL1 concentration for FH+ versus FH- subjects (Figure 3.4f). This is in accordance with the findings of Teng *et al*. [141]. Explaining the observed high LDL1 requires distinguishing between lipolysis through LPL and HL. A previous study on mechanistic modelling of the lipoprotein life cycle [21] suggests that lipolysis outside the liver by LPL mostly affects larger apoB-containing lipoproteins such as VLDL. On the other hand, HL mostly targets smaller IDL through LDL particles in the hepatic perisinusoidal space (Figure 3.5). Given the impaired binding of apoB-containing particles to LDLR on the liver, FH+ subjects can be expected to have a lower HL lipolysis and liver clearance than FH- subjects. This reduced HL lipolysis explains the accumulation of LDL1 particles [142]. Therefore, even though other LDL subfractions will increase due to a longer circulation time, accumulation of the larger LDL1 particles is especially marked. This is in agreement with the average levels of different LDL subfractions presented by FH+ and FH- subjects (Annex 2).

In conclusion, the biochemical parameters identified in this study that best discriminate between FH+ and FH- individuals are biologically plausible and provide insights into the predominant lipid pathways affected in each case.

2. Identification of different dyslipidaemic profiles among individuals by a hierarchical clustering analysis

2.1. Identification of a third class of individuals by clustering analysis

To unravel the biochemical patterns among FH+ and FH- individuals, while searching for potential biomarkers that may improve patients distinction, a hierarchical clustering of principal components (HCPC) analysis was carried out separately on each subset of the PFHS-ped. These data subsets were previously established for model training (see section 1 of this chapter), taking into account that the measures of some lipid parameters were not available for all the 211 individuals that comprise the PFHS-ped dataset. Given the fact that clustering is an unsupervised approach, the classification of individuals as FH+/FH- (variable "Class") was only considered for interpretation of clustering results and thus did not contribute to the establishment of clusters by HCPC.

Results showed that in all subsets individuals were distributed in three distinct clusters (Figure 3.6), in stark contrast to the traditional assignment of patients into two classes based on their genetic profile (FH+ and FH-). This suggests the existence of a third group of individuals characterised by a distinct lipidic pattern, based on a diverse set of biochemical parameters.

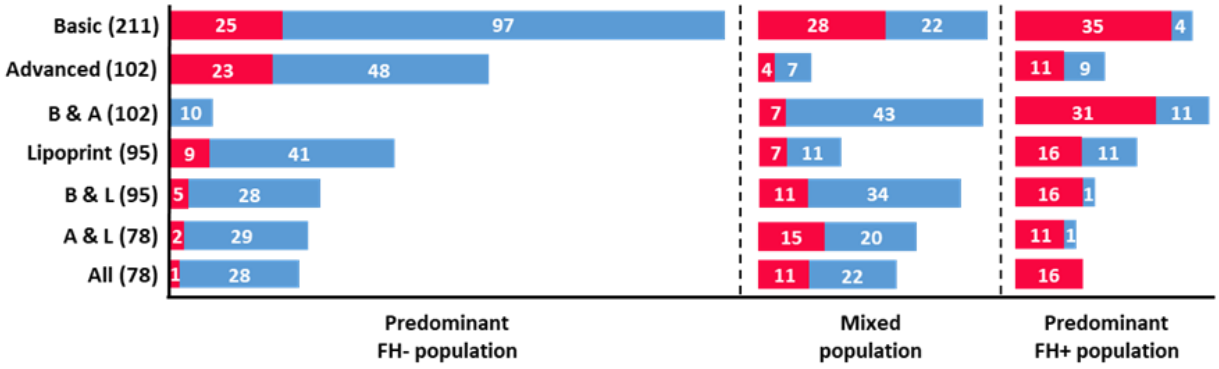


Figure 3.6. Characterization of clusters regarding the number of individuals according to their class, for each of the seven subsets. Red and blue bars correspond to FH+ and FH- populations, respectively. The number of individuals for each class is present in white. Each subset presents the total number of individuals within brackets. B & A: "Basic & Advanced"; B & L: "Basic & Lipoprint"; A & L: "Advanced & Lipoprint".

In order to obtain insights into the nature of these three groups, we began to look at the distribution of FH+ and FH- individuals among clusters. As shown in Figure 3.6, in every subset it was possible to identify a pattern in the distribution of individuals according to their class. There was always a cluster mainly constituted by FH+ patients and another cluster mostly composed by FH- individuals. In addition, a "mixed" population was present in a third cluster, including a considerable number of both FH+ and FH- individuals. In spite of the smaller

number of individuals in comparison to the other subsets, the “All” subset presented the most defined distribution of individuals among the three clusters, regarding FH+/FH- classification (Figure 3.6). After the “All” subset, the best distribution pattern was found in the subsets “Basic & Advanced”, “Basic & Lipoprint” and “Advanced & Lipoprint”. This suggests that a combination of parameters from different lipid profiles contributes to a better distinction between individuals, which is in agreement with the results obtained using a supervised ML approach considering a population with two classes (see section 1). Therefore, clustering analysis was focused on the “All” subset (called from this point as “work population”), whose distribution of individuals by clusters is shown in detail in Figure 3.7.

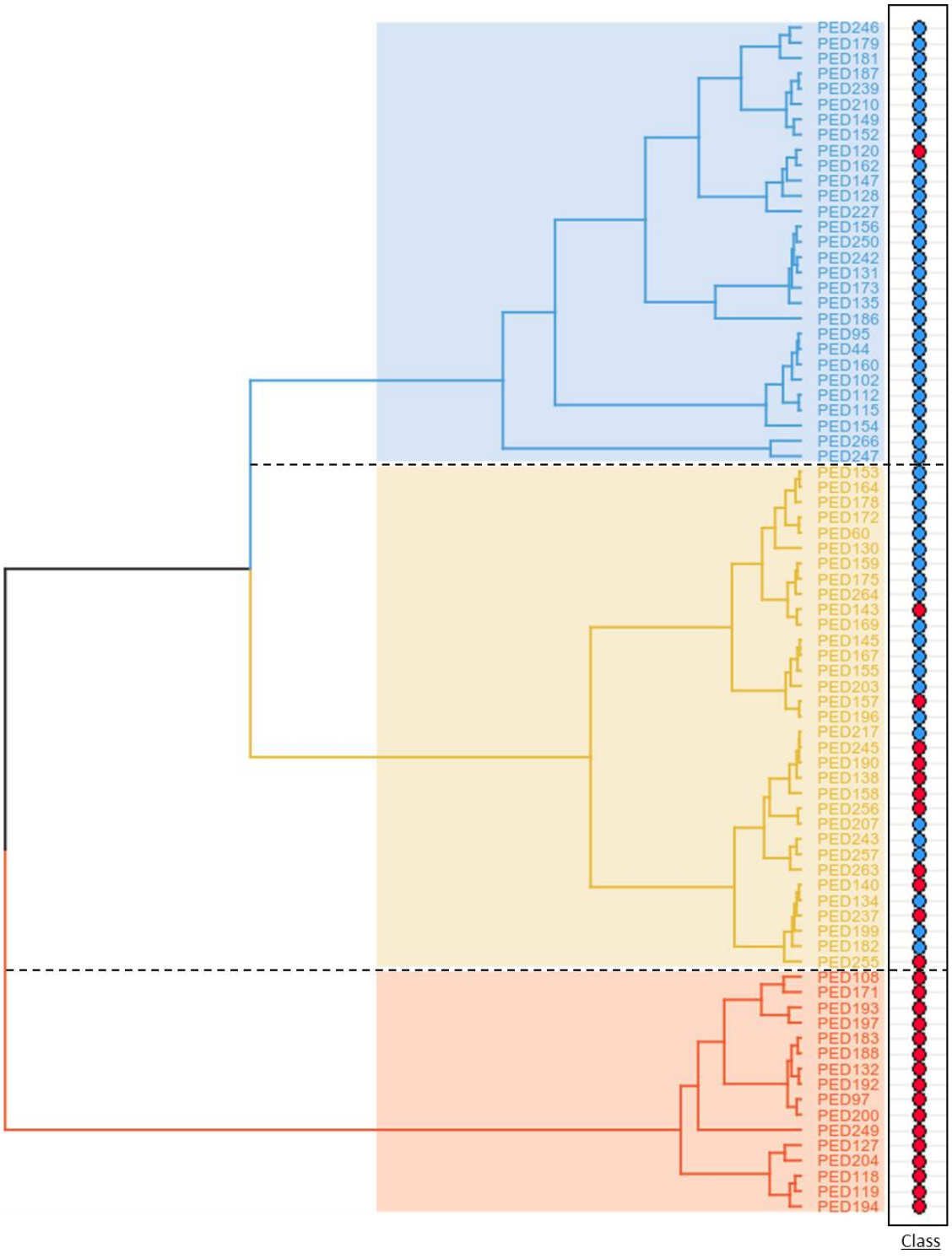


Figure 3.7. Dendrogram of the "All" subset showing the best distribution of individuals by the three clusters. Blue, yellow and red coloured clusters correspond to predominant FH-, mixed, and predominant FH+ populations, respectively. In the right panel, red and blue circles identify FH+ and FH- individuals, respectively.

As mentioned before (see methods), the HCPC analysis allowed us to acquire a detailed description of clusters regarding the contribution of both quantitative and categorical variables, individuals, and dimensions. This information is important for an accurate characterization of

clusters and to explain the distribution of individuals within each cluster, as well as to identify potential lipid profile patterns.

2.1.1. Cluster description by quantitative variables identifies metabolic pathways of interest for differentiation of FH+ and FH- classes

To understand which parameters contributed to the cluster partition, the statistical analysis carried out by HCPC (see methods) was carefully explored. This analysis took into account all the quantitative variables used by the clustering algorithm, which comprise all the biochemical parameters from PFHS-ped, BMI and age. Accordingly, the results included a ranked list of the statistically significant parameters for the distribution of individuals among three clusters (Table 3.5). The parameters were ranked according to the correlation ratio between each parameter and cluster partition, which means that parameters with the highest correlation are at the top of this list. TG/ApoB was the parameter that mostly contributed to the cluster partition, followed by VLDL, LDL-C, apoB/apoA-I and TC/HDL-C. Conversely, MIDA, HDL-C, Lp(a), VLDL/IDL and sdLDL/LDL-C were the five parameters with the smallest significant contribution to the clustering. From a total of 30 quantitative variables, only five were not considered as statistically significant for the establishment of clusters (i.e., the correlation ratio was not significantly different from zero), including apoC-II/apoC-III, apoE, LDL2, and Lipoprint measurements of HDL and sdLDL.

Table 3.5. Ranked list of the statistically significant parameters for the distribution of individuals by three clusters, under the confidence level of 95%. The parameters were ranked from the highest to the lowest significant contribution to cluster partition, according to the correlation ratio (η^2) between each parameter and cluster division.

Rank	Variables	η2	p-value
1	TG/ApoB	0.56	3.07E-14
2	VLDL	0.55	9.93E-13
3	LDL-C	0.53	2.58E-12
4	ApoB/ApoA-I	0.51	7.49E-12
5	TC/HDL-C	0.45	2.21E-11
6	IDL	0.43	1.61E-10
7	VLDL/LDL-C	0.43	1.32E-09
8	ApoC-III	0.43	1.45E-09
9	MIDB	0.42	2.71E-08
10	TC	0.41	3.62E-08
11	ApoB	0.39	8.06E-08
12	LDL1	0.35	1.95E-07
13	TG	0.34	4.31E-07
14	MIDC	0.26	1.77E-06
15	BMI	0.24	7.53E-06
16	ApoA-I	0.23	1.46E-05
17	ApoA-II	0.23	6.33E-05
18	sdLDL.Day	0.21	7.49E-05
19	ApoC-II	0.18	3.08E-04
20	Age	0.18	6.93E-04
21	MIDA	0.17	1.03E-03
22	HDL-C	0.17	1.91E-03
23	Lp(a)	0.14	4.06E-03
24	VLDL/IDL	0.13	9.73E-03
25	sdLDL/LDL-C	0.09	1.53E-02

Once again, the importance of combining parameters from different lipid profiles to identify consistent lipid patterns among individuals is shown by the presence of “Basic”, “Advanced” and “Lipoprint” parameters (Table 3.5). Another important aspect is the presence of several ratios as significant contributors to cluster partition, which highlights the relationships between different parameters in the context of lipid metabolism.

For a better characterization of each cluster and to look for potential biochemical patterns, we explored the result of student’s t-test that compared the mean of each quantitative variable in the total work population of 78 individuals (i.e., “overall mean”) with the mean of the same variable in each cluster (i.e., “mean in category”). Accordingly, Figure 3.8 shows the list of parameters whose “mean in category” was significantly different from the “overall mean”. These are the parameters that best characterise each cluster.

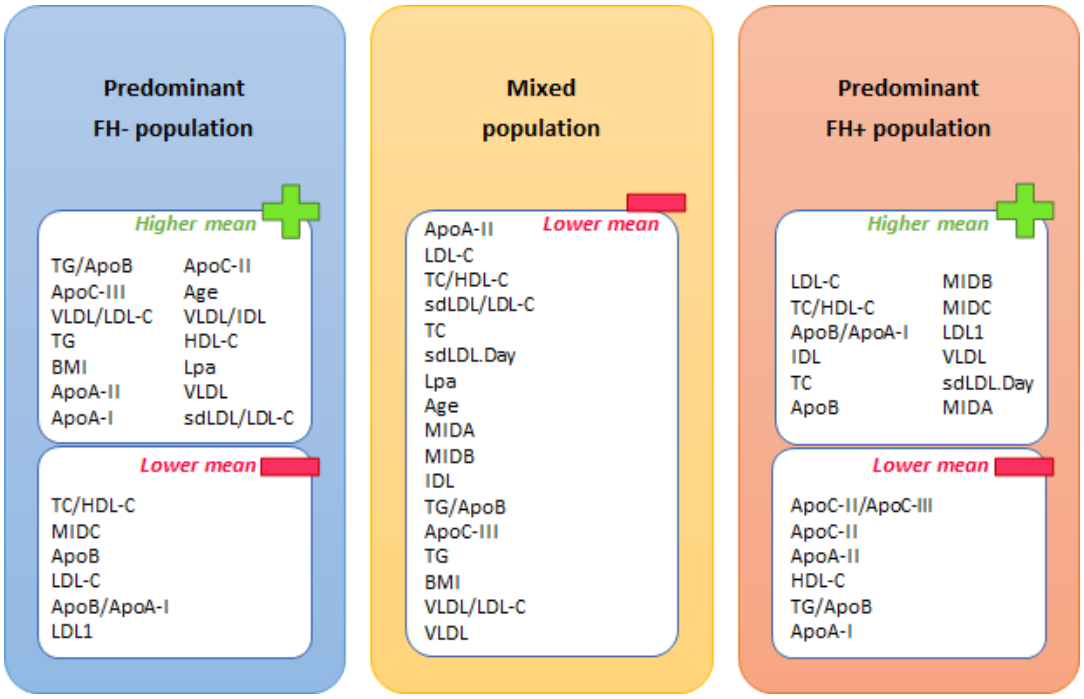


Figure 3.8. Parameters that best characterise each cluster, according to the difference between the "mean in category" of a given parameter and its "overall mean". Among the significant parameters, those that present higher/lower mean values in the cluster in comparison to the overall mean are represented in separate boxes within the cluster. The confidence level of 95% was considered for this analysis.

The cluster with a predominant FH- population is mainly characterised by higher values of parameters related to TG metabolism and lower values of parameters related to LDL and apoB metabolism. The inverse association was found in the cluster with FH+ as the predominant population. Regarding the mixed population, both TG and LDL/apoB related parameters have lower mean values than in the work population. Further details of each parameter distribution and mean trend among clusters are present in Annex 4.

2.1.2. Cluster description by categorical variables identifies molecular, biochemical, and anthropomorphic patterns among individuals

Concerning the characterization of clustering results by a set of categorical variables associated with PFHS-ped (Annex 5) and described in methods, the HPCP algorithm used a chi-squared test to measure the association between each categorical variable and cluster partition. Table 3.6 presents the variables that significantly explain the distribution of individuals by three clusters. These variables did not contribute to cluster partition, instead they were included in HCPC analysis as supplementary variables and thus only used for interpretation purposes.

Table 3.6. Categorical variables that present a statistically significant association with cluster partition results, under the confidence level of 95%.

Variables	p-value
Class	7.83E-10
Gene	3.52E-08
Activity class	1.53E-06
SB criteria	3.54E-03
BMI class	1.75E-02

As previously shown by Figure 3.7, the classification of individuals as FH+/FH- is closely associated with their distribution among clusters (Table 3.6). The affected gene in FH+ patients (*LDLR*, *APOB*, *PCSK9*) is also associated with clustering results, as well as the percentage of molecular activity that is kept by the affected allele (variable "Activity class"). These results show the relevance of genotype to patients differentiation. In addition, Table 3.6. presents that the fulfilment of SB criteria and the BMI class are also significantly associated with cluster distribution, which suggest the presence of different degrees of disease severity among individuals and the potential contribution of environmental factors (e.g., rich fat diet) for the establishment of cluster patterns, respectively.

For a more detailed description of each cluster, the association between each of them and categorical variables was statistically tested by the HCPC algorithm (see methods). Figure 3.9 presents the categories of different variables that provide a statistically significant association (positive or negative) with each cluster. Of note, none of the categorical variables presented a significant association with the mixed population cluster.

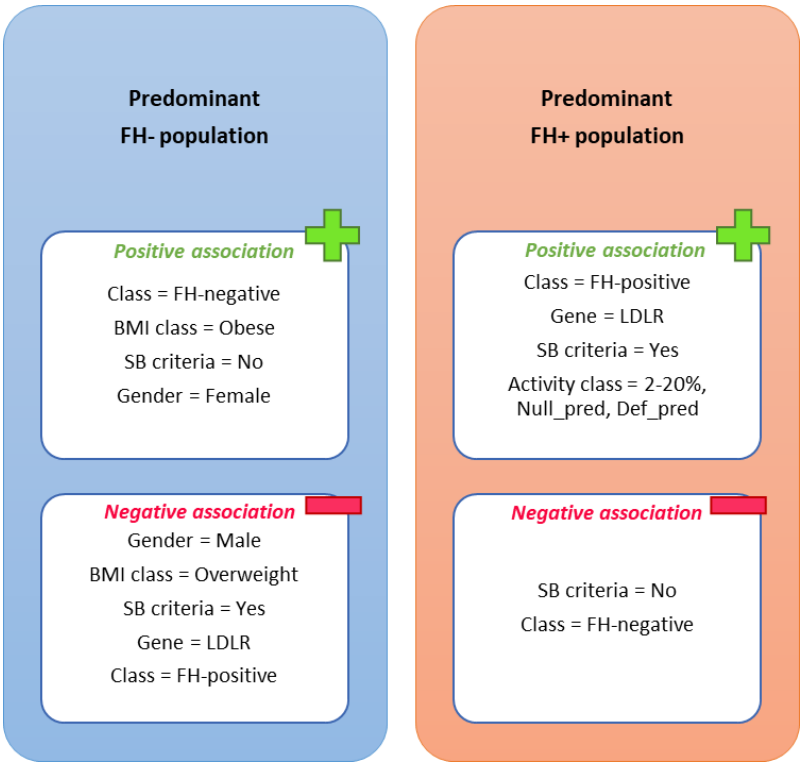


Figure 3.9. Variables categories that present a statistically significant association, positive or negative, with each of the clusters. Mixed population is not represented since no significant association was found between this cluster and categorical variables. The confidence level of 95% was considered for this analysis. The categories "Null_pred" and "Def_pred" of activity class are related to the results of *in silico* predictions for allele activity, corresponding to predicted null and defective variants, respectively.

As previously shown in this section, there is a clear pattern in the distribution of individuals among clusters according to their classification as FH+ or FH-, which is present by Table 3.6 and Figure 3.9 where variable "Class" has a statistically significant association with FH+ and predominant FH- clusters. The variable "SB criteria", regarding the fulfilment of TC and LDL-C cut-offs from SB clinical criteria, also showed a significant association with clusters. Accordingly, the predominant FH+ cluster was associated with higher levels of TC and LDL-C, compared to the predominant FH- cluster that appeared to present a milder biochemical profile. Giving the fact that *LDLR* is the most affected gene in FH [54], it is not surprising the presence of a positive association between the predominant FH+ cluster and an affected *LDLR* gene, as well as the negative association of *LDLR* with the predominant FH- cluster. In addition, the predominant FH+ cluster seemed to be characterised by an activity of 2-20% in the affected allele, besides being associated with null (less than 2% of molecular activity) and defective (2 - 80% of molecular activity) *in silico* predicted variants - i.e., "Null_pred" and "Def_pred", respectively. This suggests that the predominant FH+ cluster is associated with a more severe genotype, with the significant presence of variants keeping a molecular activity not higher than

20% in comparison to wild type. Regarding other variables like gender and BMI class, the predominant FH- cluster appeared to be characterized by the presence of obese girls, in contrast to the negative association with overweight boys. Still, this association of individual distribution among clusters with gender should be considered with caution, as explained forward (see discussion).

In pursuit of a clear visualisation of clusters characterization by categorical variables, the distribution of these variables was compared with the distribution of individuals among clusters, as shown in Figure 3.10.

Class	SB criteria	Gender	Lipoprint profile	Colour
FH+	Yes	Female	B	
FH-	No	Male	A	

BMI class	Colour
Severe thinness	
Thinness	
Normal	
Overweight	
Obese	

Activity class	Colour
Null	
2-20%	
20-40%	
40-65%	
Null_pred	
Def_pred	

LDL-C score	Colour
≥0.9	
0.9 – 0.7	
0.7 – 0.5	
<0.5	

Gene	Colour
LDLR	
APOB	
PCSK9	

Figure 3.10. Distribution of categorical variables within the classification dendrogram. White circles on "Activity class" and "Gene" correspond to FH- individuals, while on "LDL-C score" they represent the individuals whose score was not performed.
(Continued in the next page)

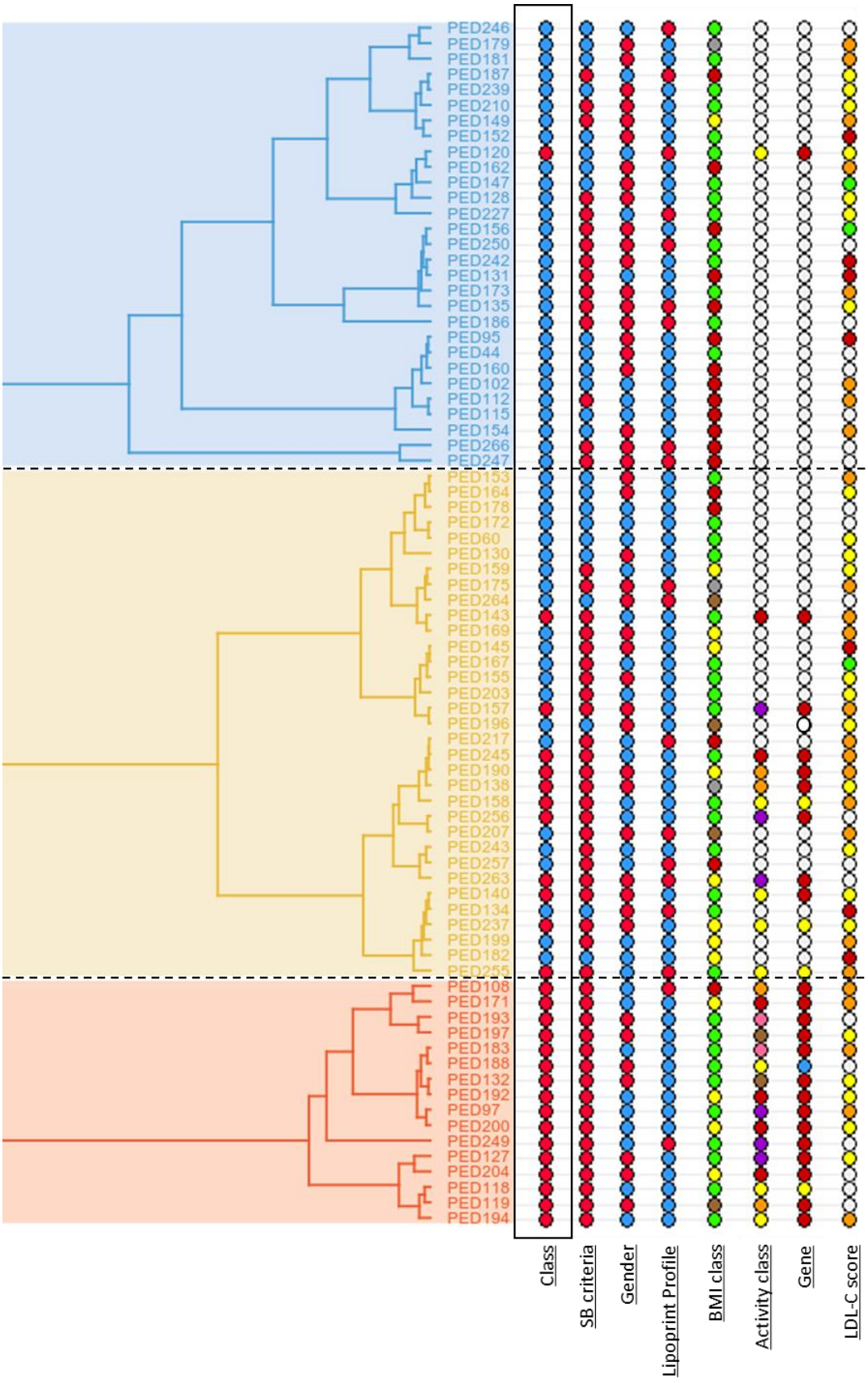


Figure 3.10. **(Continued from previous page)** Distributive pattern of the categorical variables according to position of individuals within the dendrogram. White circles on "Activity class" and "Gene" correspond to FH- individuals, while on "LDL-C score" they represent the individuals whose score was not performed.

Apart from the already recognized pattern of variable “Class”, some less specific patterns could be identified with a careful observation of Figure 3.10 considering the distribution of other categorical variables. Still, a clear distinction among individuals was not achieved in the mixed cluster.

The most severe genetic variants corresponding to lower percentages of molecular activity, comparatively to the wild type allele, seems to agglomerate in the predominant FH+ cluster and in the first individuals of the mixed cluster. This can indicate a decreasing ruler of disease severity once we move in direction of the predominant FH- population. Still, there are some individuals on the right branch of the mixed cluster (i.e., PED263, PED256, PED245, PED157, PED143) that present extremely low levels of activity in the affected allele. Considering that more severe variants translate in higher levels of TC and LDL-C, other lipid parameters that were not possible to identify may be responsible for setting these FH+ individuals closer to a FH- profile, which could explain their assignment to the mixed cluster. Conversely, PED120 is the only FH+ belonging to the predominant FH- cluster, which is due to the presence of a pathogenic variant of mild effect (c.1216C>T, p.Arg406Trp) and that translates in a milder phenotype in comparison to all the other FH+ individuals [49].

Predominant FH+ cluster is totally composed of individuals that fulfil TC and LDL-C cut-offs from SB criteria, which emphasises the idea that this cluster presents the most severe profiles. Accordingly, the number of individuals fulfilling these cut-offs decreases as we move forward through the mixed cluster and, even more, within the predominant FH- cluster. This means that subjects from predominant FH+ cluster present higher levels of TC and LDL-C in relation to the other clusters. In contrast, the LDL-C score appears to be higher once we move through the mixed cluster and, especially, within the predominant FH- cluster. This means that this last cluster presents a higher number of individuals having a stronger polygenic contribution for their biochemical profiles, namely LDL-C levels [36].

Regarding gender, there is a higher number of females once we move in the direction of the predominant FH- cluster. The same trend is also observed for the category “obese” of the variable “BMI class”, which suggests that within the PFHS-ped dataset a FH- profile is easier to find among girls suffering from obesity. This finding was also observed while testing the statistical association between clusters and each of the categorical variables (Figure 3.9). Still, gender did not show a significant association with the cluster partition (Table 3.6) and the predominant FH- cluster was the only cluster that presented an unbalanced number of individuals from one gender in relation to the other.

As mentioned before, the distribution pattern of "Gene" variable is in agreement with the classification of individuals as FH+ or FH- and, since *LDLR* variants are identified in a big majority of FH cases, this explains the considerable presence of *LDLR* variants in the predominant FH+ cluster and, in less extent, within the mixed cluster. In addition, there are four *APOB* variants among these two clusters, three of them in the mixed cluster. *APOB* variants are known to produce milder phenotypes in comparison with *LDLR* variants [54]. This highlights the differences in disease severity between clusters.

In relation to "Lipoprint profile", the predominant FH+ cluster seems to present less individuals of profile B in relation to the other clusters. As explained before (see methods), the Lipoprint profile is obtained by measuring the concentration of sdLDL particles in serum, which correspond to the most atherogenic fraction of LDL. Then, a profile B is related to a higher cardiovascular risk, since a higher concentration of these particles are present, in comparison to a subject with a profile A [16], [143]. This suggests a valuable insight regarding the potential influence of an affected HL activity and consequent change in proportions of LDL subfractions on disease severity, as explained forward in Chapter 4.

2.1.3. Cluster description by individuals emphasises the presence of different dyslipidaemic profiles besides the categorization as FH+ and FH-

In addition to the statistical analysis that tested the variable contribution for clusters, the HCPC algorithm has allowed us to assess which individuals best describe and/or are more specific of each cluster, by measuring the distance between each individual and the gravity centre of each cluster. Table 3.7 presents the paragons, which are the individuals considered to best characterise each of the clusters, since they are the closest individuals to the cluster centre.

Table 3.7. Individuals with the shortest distance to the centre of the cluster they belong.

Predominant FH- cluster		**Mixed cluster**		**Predominant FH+ cluster**	
Patient	*Distance*	*Patient*	*Distance*	*Patient*	*Distance*
PED131	0.83	PED159	0.85	PED192	2.11
PED187	1.03	PED140	1.16	PED132	2.27
PED239	1.93	PED237	1.77	PED188	2.56
PED242	2.12	PED138	2.01	PED171	2.62
PED162	2.31	PED217	2.02	PED193	2.64

As shown in Table 3.7, PED131, PED159 and PED192 are the individuals that best characterise predominant FH- cluster, mixed cluster and predominant FH+ cluster, respectively, since they are the ones with the shortest distance to each of the cluster's centres. From these three subjects,

only PED192 is FH+ and PED131 is the one presenting the higher BMI and polygenic score. This is in agreement with previous results in this chapter that suggested a more severe dyslipidaemic profile for patients of predominant FH+ cluster, while a higher polygenic contribution and BMI might be associated with the predominant FH- cluster. Further, the individuals that best represent the predominant FH+ cluster are mostly carriers of more severe FH-associated gene variants, with a normal BMI and a moderate polygenic contribution. On the opposite, predominant FH- cluster are best described by FH- individuals that in average present a high polygenic score and a higher BMI than normally expected, with some of them also presenting milder TC and LDL-C levels in comparison to individuals from other clusters. The mixed cluster appears to represent a mixed phenotype with characteristics of both FH+ and FH- profiles, for the following reasons: there is only one FH+ among the list of representative individuals, whose pathogenic variant in *APOB* gene is considered milder than the *LDLR* variants associated to the predominant FH+ cluster [54]; the pattern of polygenic contribution seems to be similar to those of predominant FH+ cluster, whilst the polygenic scores of PED188 and PED193 were not performed; the BMI pattern is worse in comparison to the predominant FH+ cluster but milder than the one found in the predominant FH- cluster; TC and LDL-C levels are higher than those of the predominant FH- cluster. Still, the low number of individuals involved in this analysis (five individuals as best representatives of each cluster) should be taken into consideration for further discussion around the mixed population.

On the other hand, Table 3.8 presents the individuals that are more distant to the centre of other clusters and that can be considered the most specific individuals of their own cluster.

Table 3.8. Individuals with the longest distance to the centre of other clusters, according to the cluster they belong.

Predominant FH- cluster		**Mixed cluster**		**Predominant FH+ cluster**	
Patient	*Distance*	*Patient*	*Distance*	*Patient*	*Distance*
PED247	11.59	PED130	6.89	PED249	9.41
PED186	10.43	PED203	6.59	PED97	7.52
PED266	9.60	PED243	6.30	PED183	7.11
PED227	8.08	PED143	5.73	PED197	6.85
PED128	6.91	PED257	5.66	PED192	6.78

As shown in Table 3.8, PED247, PED130 and PED249 are the most specific individuals of predominant FH- cluster, mixed cluster, and predominant FH+ cluster, respectively, since they present the longest distances to the centre of the other clusters. PED249 presents a normal BMI and high levels of TC and LDL-C, besides carrying a null variant in the *LDLR* gene, which is in agreement with the pattern already associated with the predominant FH+ cluster. Indeed, the

most specific individuals of this cluster have mostly high values for TC and LDL-C and present in average a moderate to high polygenic contribution and a normal BMI, besides of three of them being carriers of more severe gene variants. In other hand, the most specific individuals of the predominant FH- cluster mostly present TC and LDL-C levels that fulfil the cut-offs from SB criteria, and a pro-atherogenic profile regarding the concentration of sdLDL measured by Lipoprint assay. In addition, two of these individuals were considered obese. In relation to the mixed cluster, the most specific individuals mostly present TC and LDL-C levels above SB cut-offs, besides of a normal BMI and a moderate polygenic contribution. PED143 is the only FH+ individual of the mixed cluster, with a gene variant associated with 2-20% of molecular activity in the affected allele, which is accompanied by a very high polygenic score. Still in the same cluster, PED257 is the only patient with a high BMI and a pro-atherogenic Lipoprint profile. As mentioned before, we should be careful while taking conclusions from these results, considering the low number of individuals involved.

PED192 belongs simultaneously to the lists of Table 3.7 and Table 3.8 regarding the predominant FH+ cluster, which means that this individual is not only among those who best characterise this cluster but also one of the most specific of its members. This patient presents high levels of LDL-C and TC, a high ratio apoB/apoA-I, besides carrying a pathogenic variant associated with 2-20% of molecular activity in the affected allele. This pattern is aligned with previous findings mentioned in this section, regarding the predominant FH+ cluster.

2.1.4. Cluster characterization by dimensions highlights the presence of different metabolic patterns among individuals

To acquire a more detailed cluster description, besides the characterization by variables and individuals, the clustering algorithm assessed the PCA dimensions that best explain distribution of individuals among clusters. For this the HCPC algorithm has tested the association between individual coordinates within each cluster and the different dimensions (also known as principal components or axes in the context of PCA results), which translates in the identification of the dimensions that best describe each cluster. Table 3.9 shows the dimensions where individuals present the statistically significant stronger or weaker coordinates, for each of the three clusters.

Table 3.9. Dimensions that present statistically significant associations with individual coordinates within each cluster, under the confidence level of 95%.

Predominant FH- cluster		
Dimensions	*p-value*	*Association*
PC2	5.22E-08	stronger
PC4	0.005896	stronger
PC1	0.000114	weaker
Mixed cluster		
Dimensions	*p-value*	*Association*
PC3	0.0394	weaker
PC4	2.17E-05	weaker
PC2	1.79E-06	weaker
Predominant FH+ cluster		
Dimensions	*p-value*	*Association*
PC1	7.11E-12	stronger

As depicted in Table 3.9, individuals from the predominant FH- cluster have a stronger association with PC2 and PC4 in comparison to other clusters, while this association is weaker for PC1. Conversely, individuals from the predominant FH+ cluster present a stronger association with PC1. Concerning the mixed cluster, these individuals have a weaker association with PC3, PC4 and PC2 in relation to individuals of other clusters.

Looking through the PCA results, it is possible to access which variables most contribute to each dimension. Crossing this information with the dimensions that are significantly associated with each cluster (Table 3.9) allowed to consolidate the description of clusters. Giving the fact that only the first five dimensions were used for the HCPC analysis (see methods), and from these only four presented a significant correlation between at least one of the three clusters (Table 3.9), the variable contribution by dimension was analysed only for PC1, PC2, PC3 and PC4 (Figure 3.11).

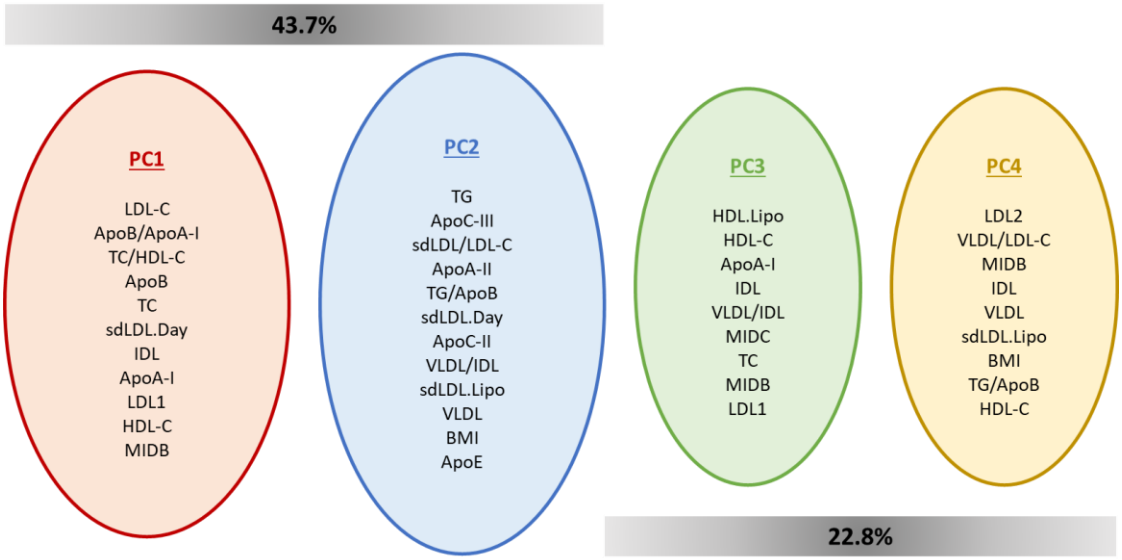


Figure 3.11. Variables that most contribute for each of the dimensions that previously have shown to be significantly correlated to clusters. PC1 and PC2 together contribute to explain 43.7% of variance in the dataset, while PC3 and PC4 together explain 22.8% of data variance.

The variables that contribute most for PC1 were mainly linked to LDL/apoB metabolism, while PC2 were mostly associated with variables related to TG metabolism and ratios that establish interactions between TG and LDL/apoB pathways. This emphasises the relation that was already established between the predominant FH+ cluster (showing a stronger association to PC1) and LDL/apoB metabolism. Conversely, the predominant FH- cluster, presenting a weaker association to PC1 and a stronger association to PC2, was previously associated with TG metabolism. This cluster also presented a strong correlation with PC4, whose main contribution came from variables associated to the different LDL fractions, besides other parameters like BMI, TG/apoB and HDL-C. In contrast, PC4 presented a weaker association with the mixed cluster, which was also reported for PC2 and PC3 that were both related to parameters involved in TG metabolism and LDL pathway/reverse cholesterol transport, respectively. This translates to lower values of these parameters for individuals from the mixed cluster in comparison to individuals from other clusters.

Given the fact that PC1 and PC2 explain more than 40% of data variance, which make them the most informative dimensions, the distribution of individuals within each cluster was plotted for PC1 and PC2 dimensions (Figure 3.12).

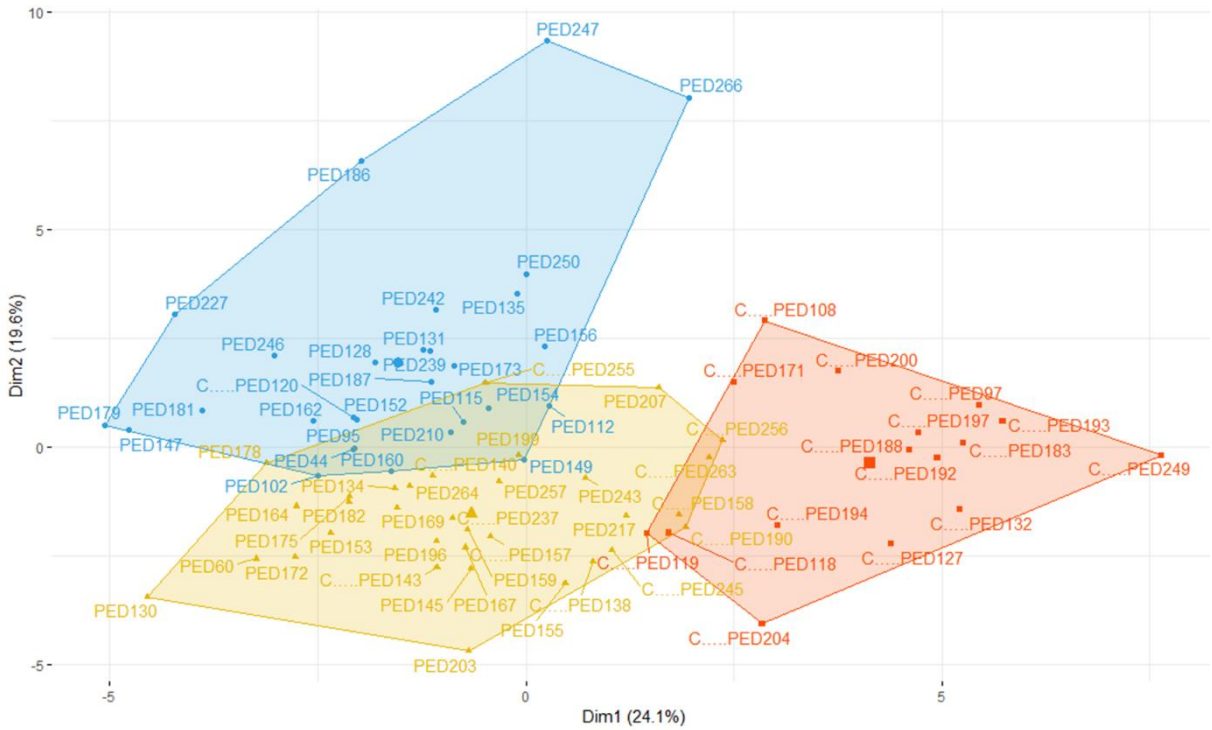


Figure 3.12. Cluster map showing the distribution of individuals within each cluster for PC1 (also known as Dim1) and PC2 (also known as Dim2), which correspond to the dimensions that best explain data variance. FH+ individuals are identified with a "C" before their ID. Predominant FH- cluster in blue, mixed cluster in yellow, and predominant FH+ cluster in red.

Figure 3.12 clearly shows the strong correlation of the predominant FH+ cluster with PC1, since the coordinates of individuals are all positive for this dimension. Conversely, most of the individual coordinates within the predominant FH- cluster are positives for PC2 and negatives for PC1. Regarding the mixed cluster, a considerable part of these individuals presents negative coordinates for both PC1 and PC2 dimensions.

2.2. Predicted class assignment using Imp_B model suggests the presence of borderline individuals

As explained before, contrary to the predominant FH+ and predominant FH- clusters, the mixed cluster could not be clearly characterised since it was not possible to identify specific biochemical or clinical patterns. Then, for a better characterization of this cluster, a previously trained model (Imp_B, see Table 3.2) was applied to the 78 individuals of the work population, letting to predict classification with the associated probability of each individual belonging to FH+ and FH- class. This allowed the identification of individuals whose classification can be potentially dubious, especially because of milder or severer biochemical profiles for FH+ and FH- subjects, respectively, which might be close to the borderline between these two classes.

After acquiring the probabilities associated with the predicted classifications, the difference between the probability of being FH+ and the probability of being FH- (named as Δprob) was

measured for each individual. For a better visualisation and results interpretation, the set of values of Δprob were grouped into four categories: very ambiguous (Δprob <0.25), ambiguous (0.25 ≥ Δprob < 0.5), reasonable (0.5 ≥ Δprob < 0.75), clear (Δprob ≥ 0.75). Accordingly, a high Δprob corresponds to a clear classification, while a low Δprob is associated with an ambiguous classification. Figure 3.13 shows the distribution of Δprob among individuals within clusters. The predicted class assignment, associated probabilities, Δprob and correspondent categories are available for all individuals of the work population in Annex 6.

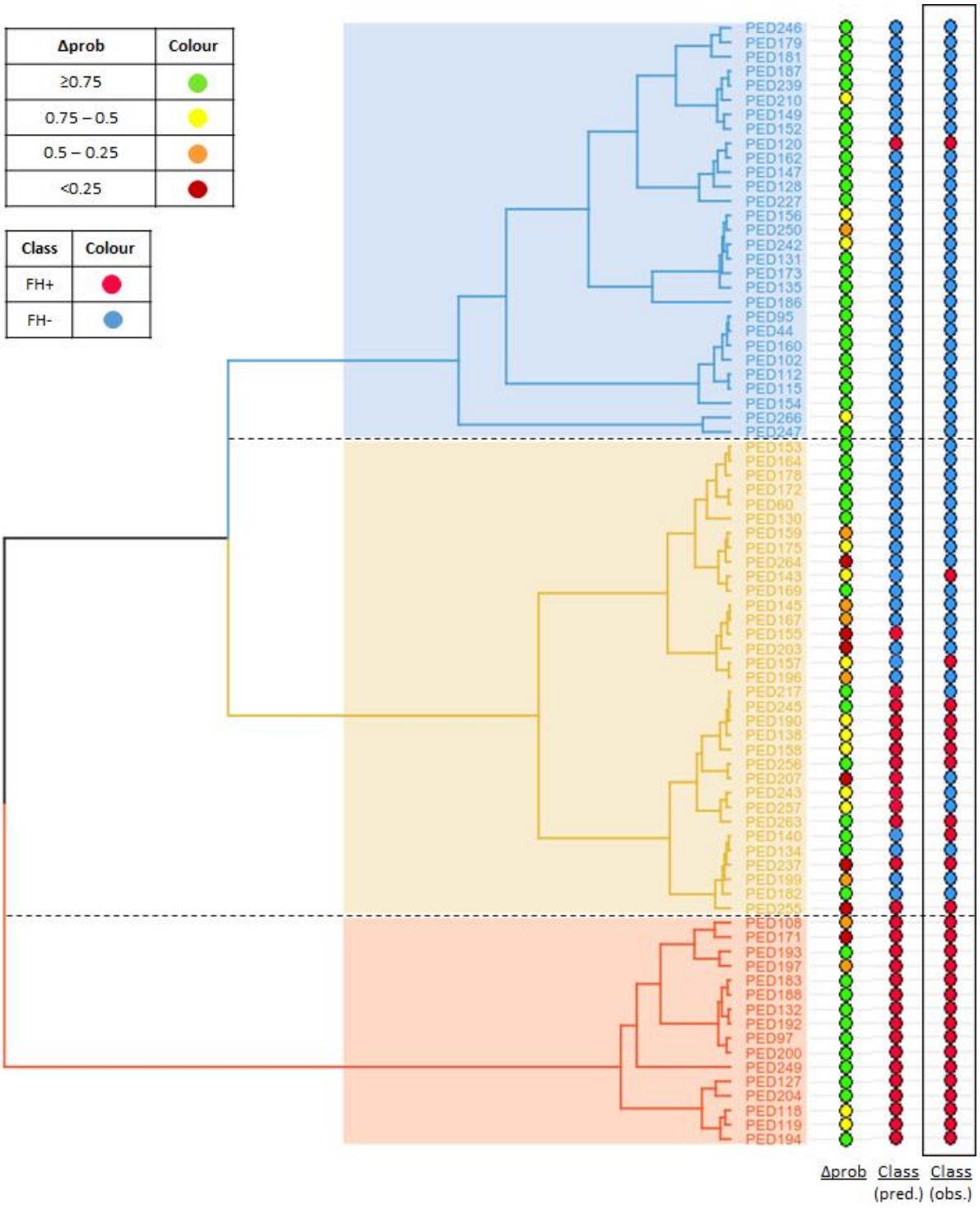


Figure 3.13. Distribution of Δprob across the 78 individuals of the work population, ordered according to cluster position. The value of Δprob is the difference between the probability of being classified as FH+ and the probability of being classified as FH- according to the predictions obtained using the Imp_B model. Both predicted and observed classifications are also shown in the figure. The colour label corresponding to the different categories of Δprob and classification (FH+/FH-) are present in the upper left corner.

As seen in Figure 3.13, most individuals considered to have an ambiguous classification, according to the predictions acquired with the "Imp_B" model, are present in the mixed cluster. This is in agreement with the hypothesis of this cluster representing individuals with a mixed

phenotype, as previously mentioned in this chapter. Then, since the mixed cluster is composed of both FH+ and FH- subjects, the FH+ individuals of this cluster seem to present a biochemical profile milder than individuals of the predominant FH+ cluster, while the FH- subjects appear to have a more severe profile than individuals of the predominant FH- cluster. We should take into account that dyslipidaemia is a complex disease and that the final phenotype presented by an affected individual results from the interaction of different genetic and environmental factors, including both monogenic and polygenic contributions in the hallmark FH genes and/or in other lipid-related genes, epigenetic factors, diet and lifestyle (e.g., physical activity or smoking habits) [10], [45].

The clustering analysis performed on the PFHS-ped dataset allowed the identification of different metabolic profiles between FH+ and FH- individuals, including the association of each class with LDL/apoB and TG metabolism, respectively, as already suggested by modelling analysis. In addition, the application of this explorative approach resulted in the classification of individuals into three groups, with two of them being clearly characterised by a FH+ and FH- profile, respectively, while a third group was shown to comprise individuals with mixed phenotype representing the biological complexity of dyslipidaemia. The characterization of each of these groups of individuals allowed the identification of potential biomarkers that may be useful to future genetic studies and/or new classification models involving larger datasets.

3. Creation of a new lipid knowledge base directed to dyslipidaemia

3.1. Defining a list of target genes and collecting different levels of gene information

For a better understanding of the dyslipidaemia biological context, a list of genes of interest was established and associated information (i.e., gene expression data, associated GWAS traits and GO terms) was collected from public databases. The integration of all this information into a new knowledge base may allow us to look for potential expression, phenotype and/or functional patterns among these genes, promoting the identification of new biomarkers and a better discrimination between dyslipidaemic individuals. Therefore, within a perspective of integrative analysis, a set of genes/proteins, metabolites and metabolic pathways were selected as keywords (Table 3.10) to search for target metabolic pathways in the Wikipathways platform. Some of these keywords represent potential biomarkers for the distinction of different dyslipidaemic patients, considering the previous results achieved by ML-based methods. Briefly, the parameters related to TG and LDL/apoB metabolism were shown to contribute for a better distinction between individuals, according to both modelling and clustering analysis. The best ranking models were mostly composed of a set of biochemical parameters that included TG, LDL-C, apoB, apoA-I, apoC-III and LDL1. The first four of these parameters were also present in the top five of statistically significant contributors for cluster partition, which also included VLDL. The results have also shown that lipoprotein metabolism is one of the major pathways involved in the dyslipidaemia biological context. In addition, other terms were considered as keywords, including lipid metabolism, hypercholesterolaemia and atherosclerosis (Table 3.10), based on the literature review.

Table 3.10. Biological entities used as keywords in Wikipathways search tool, which allowed to identify a set of metabolic pathways of interest. In addition to genes, metabolites and pathways selected from previous results, other keywords were included considering the literature review.

Keyword	Class	Keyword	Class
Lipid metabolism	Pathway	VLDLR	Gene
Lipoprotein metabolism	Pathway	APOB	Gene
Triglycerides	Metabolite	LDL1	Metabolite
LDLR	Gene	IDL	Metabolite
APOA1	Gene	Hypercholesterolaemia	Disease
APOC3	Gene	Atherosclerosis	Disease

Afterwards, 14 metabolic pathways were selected (Table 3.11) and their genes were compiled in a single list of 466 genes – called target genes.

Table 3.11. Metabolic pathways that were selected after searching in Wikipathways for previously identified biological entities of interest. The number of genes out of the 466 that belong to each of the pathways is presented, with some genes being associated to more than one pathway.

Wikipathways ID	Metabolic Pathways	Genes
WP3926	ApoE and miR-146 in inflammation and atherosclerosis	9
WP3601	Composition of lipid particles	9
WP2764	Lipid digestion, mobilization and transport	70
WP4051	Lipid particle organization	6
WP4129	Plasma lipoprotein assembly, remodeling and clearance	70
WP1885	Platelet homeostasis	81
WP2878	PPAR alpha	25
WP3942	PPAR signaling	68
WP2797	Regulation of lipid metabolism by PPAR alpha	121
WP2011	SREBF and miR-33 in cholesterol and lipid homeostasis	17
WP1982	SREBP signaling	70
WP430	Statin pathway	31
WP4131	Triglycerides metabolism	37
WP1533	Vitamin B12 metabolism	54

miR: micro-RNA; PPAR: peroxisome proliferator-activated receptor (family); SREBF: sterol regulatory element-binding transcription factor (family); SREBP: sterol regulatory element-binding protein (family)

Of note, the pathway "Lipid digestion, mobilization and transport" (Table 3.11) was divided in five different pathways according to the most recent version of Reactome database [144], as follows: intestinal lipid absorption (R-HSA-8963678); digestion of dietary lipid (R-HSA-192456); triglyceride catabolism (R-HSA-163560); lipid particle organization (R-HSA-8964572); plasma lipoprotein assembly, remodeling, and clearance (R-HSA-174824).

Regarding gene expression data, for categorization of target genes according to their expression levels, cut-offs were established (see methods) for the tissues of interest, liver and small intestine given their main role in lipid metabolism, and for transcriptome (Table 3.12). The gene expression values for the transcriptome comprise median values by gene and tissue obtained using a GTEx dataset of 56202 genes and 51 tissues (see methods), representing an estimation of the median expression of each gene in the whole human organism.

Table 3.12. Gene expression categories according to the established cut-offs and the number of genes out of the 466 belonging to each category in tissues of interest and transcriptome. TPM: transcript per million

Gene expression		Number of genes		
Categories	**Cut-offs (TPM)**	**Liver**	**Small Intestine**	**Transcriptome**
Null	<0.1	42	31	6
Low	0.1 – 1	51	48	25
Moderate	1 – 10	136	110	111
Moderate-to-high	10 – 100	184	250	279
High	100 – 1000	37	27	43
Top (very high)	≥1000	16	0	2

Considering the number of genes by expression category, Table 3.12 shows that the majority of target genes belong to moderate and moderate-to-high categories, including several genes already known to be related to dyslipidaemia such as *LDLR*, *PCSK9* or *LIPA*. The coding genes of apolipoproteins, important players in lipid metabolism and dyslipidaemia, are associated with expression levels that mostly vary between moderate-to-high, high, and top levels, at least for one of the tissues of interest. In comparison with transcriptome, there is a considerable number of genes whose expression level is lower in liver and small intestine. Still, the expression pattern of the majority of genes is more similar between small intestine and transcriptome, rather than between liver and transcriptome. Comparing the two tissues of interest, there is a considerable number of genes with a lower expression in the liver than in the small intestine. The complete list of gene expression categories, in tissues of interest and transcriptome, for all target genes can be found in Annex 7.

Considering that the last report of the NHGRI-EBI GWAS Catalog refers to the presence of 5687 GWAS studies that comprises 71673 gene variant - trait associations [120], the gathering of GWAS information for target list took into account a set of keywords related with lipid metabolism and dyslipidaemia to search for GWAS traits of interest (see methods). Accordingly, these keywords translated into nine traits, including lipoprotein measurement, HDL-C, LDL-C, TC, VLDL-C, hypertriglyceridemia, coronary artery disease, cardiovascular disease, and atherosclerosis. All the nine traits were represented in the target gene list, while 96 out of the 466 target genes were associated with at least one of these traits. Figure 3.14 shows the number of target genes associated with each trait, taking into account that some genes were associated with multiple traits (maximum reported of six associated traits per gene). The full list of traits associated with each of the 96 genes is present in Annex 7.

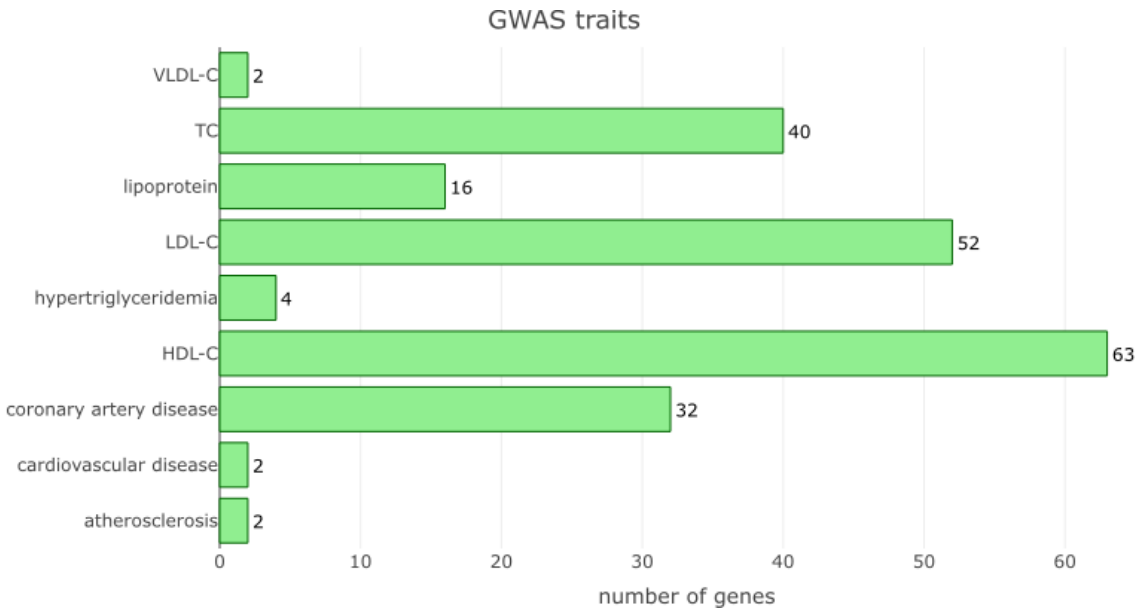


Figure 3.14. GWAS traits associated with target genes, including the number of genes related to each of the traits.

Considering the metabolic pathways selected for the establishment of target genes list (Table 3.11), it was not surprising that the GWAS traits more frequent among target genes were HDL-C, LDL-C and TC, whose central role in cholesterol and lipoprotein metabolism is already known [9], [24]. Accordingly, each of these traits were associated with at least 40 out of the 466 target genes. These GWAS traits correspond to common measured lipid parameters in clinical practice, with TC and LDL-C being part of several clinical criteria for FH (including the SB criteria). Other traits with considerable representation in the universe of target genes were coronary artery disease, which represents one of the consequences of an unbalanced lipid metabolism, and lipoprotein measurement, which stands for a more general term comprising other lipoproteins besides HDL, LDL or VLDL. The least represented GWAS traits among target genes were hypertriglyceridemia (i.e., high levels of TG), VLDL-C, cardiovascular disease, and atherosclerosis. Considering that one of the previously selected pathways was "Triglycerides metabolism" (Table 3.11) and that this pathway was identified as a potential contributor for an improved distinction between dyslipidaemic individuals (section 1 and 2 of this chapter), a higher number of target genes was expected to be associated with hypertriglyceridemia. Still, we should take into account that the number of GWAS studies is limited for some traits, which retrains the number of gene associations found until the moment. In addition, for traits like cardiovascular disease and atherosclerosis, the low representativity among target genes may be explained for the fact that these are considerably complex diseases and much more genes are involved besides the ones within target list.

In relation to functional information, a list of GO terms associated with target genes was collected for the GO domains "biological process" (BP) and "molecular function" (MF), resulting in a total of 10183 and 3909 GO terms, respectively. A manual selection considering the scope of this study and GO hierarchical relations between terms allowed to restring these lists to smaller sets of terms, yet representative of target genes (see methods). For both MF and BP domains, a group of lipid-specific GO terms was selected, comprising five (Table 3.13) and 11 (Table 3.14) parent terms, respectively. An additional selection for "Other GO terms", not lipid-specific, resulted in a set of six parent terms for each GO domain (Table 3.15 and Table 3.16). According to the GO hierarchical graph, the designation "parent" terms refers to the nodes closer to the graph's root, while "child" terms are those closer to the leaf nodes. In this study, child terms were grouped under the correspondent parent term, considering the universe of selected GO terms.

Table 3.13. List of lipid-specific parent GO terms in MF functional domain and the associated target genes. Ngc: number of genes associated with child terms of each parent term.

ID	GO term name	Genes	Ngc
GO:0016298	lipase activity	PNLIP, LPL, MGLL, LIPC, PNLIPRP2, LIPH, LIPE, LIPA, PNLIPRP3, LIPI, LIPG	23
GO:0008289	lipid binding	FABP9, FABP7, FABP6, FABP12, APOA1, FABP2, SCP2, APOE, AP2A2, PPARA, INSIG1, GPIHBP1, INSIG2, PLTP, APOA4, HDLBP, SCAP, CETP, FABP1, FABP5, CD36, PLIN1, APOA5, LIPF, PLA2G4A, MTTP, PDIA2, ALB, DBI, APOC3, APOC2, AP2M1, PPARD, FABP3, FABP4, APOA2	40
GO:0005319	lipid transporter activity	APOA2, NPC1, MTTP, ABCA1, NPC1L1, APOA1, APOC4, APOA4, APOF, APOE, APOB	23
GO:0071813	lipoprotein particle binding	LPL, GPIHBP1, APOE, CD36, APOA1	12
GO:0070325	lipoprotein particle receptor binding	APOA5, LRP1	13

Table 3.14. List of lipid-specific parent GO terms in BP functional domain and the associated target genes. Ngc: number of genes associated with child terms of each parent term. **(Continued in the next page)**

ID	GO term name	Genes	Ngc
GO:0006637	acyl-CoA metabolic process	ACSL6, DBI, GPAM	13
GO:0044241	lipid digestion	PNLIP, CEL, CLPS, PNLIPRP2	12
GO:0055088	lipid homeostasis	PNPLA5, NR1H4, ACOX1, PNPLA4, NR1H2, CETP, ABCB4, NR1H3, PPARG, ABHD5, ACOX3, ACOX2, ANGPTL3, ACACA, APOE, APOA4	55
GO:0010876	lipid localization	PPARA, SREBF1, CPT1A	85

ID	GO term name	Genes	Ngc
GO:0006629	lipid metabolic process	CPT2, ACAA1, PRKAG1, CETP, PNPLA5, LSS, FASN, HMGCS1, LPA, LPL, ANGPTL4, LIPC, SCAP, SREBF1, CEL, ACSL6, PNLIP, SLC27A6, PLTP, MOGAT3, SLC27A5, PAFAH2, NR1H3, APOE, NPC1L1, PCK1, LIPK, HSPG2, LIPH, HMGCS2, SLC27A1, ABHD5, VLDLR, LPIN3, FABP6, HMGCR, GPAT2, LIPF, APOA1, CYP7A1, CYP4A11, PPARA, MBTPS1, PCSK9, FDFT1, PPARG, ACOX1, ACSL4, MTTP, CYP51A1, LRP1, ACACA, MVD, APOF, PLA2G4A, PRKAA1, FDPS, SLC27A2, CUBN, LCAT, APOBR, NR1H4, PLIN1, PNLIPRP3, LDLR, PNLIPRP2, LIPJ, LRP2, ACADL, ABCA1, SLC27A4, PRKAB2, ACSBG2, FABP5, INSIG1, PRKAG3, APOC1, SCD, ACSL1, LIPA, ABCB4, NPC1, CPT1A, ACADM, APOB, IDI1, CIDEA, ACSL5, ACOX2, MOGAT1, CPT1B, PRKAG2, LDLRAP1, PRKAA2, SOAT2, INSIG2, LPIN1, PRKAB1, SULT2A1, ACSL3, LIPI, MGLL, ACLY, SREBF2, ACSBG1, MOGAT2, NCEH1, PTPN11, LIPE, GPAM, LIPM, CPT1C, CYP1A1, APOC3, EHHADH, LRP8, LIPN, FADS2, CD36, APOC2, HDLBP, LIPG, CLPS, ANGPTL3, PNLIPRP1, APOC4, PPARD, CYP27A1, PNPLA4, LPIN2, ANGPTL8, SOAT1, DGAT2, NR1H2, MBTPS2, ACOX3	210
GO:0042157	lipoprotein metabolic process	APOA1, APOE, APOA4, OLR1, APOC1, APOA5, APOA2, ABCA1, PPARA, MTTP, LRP1, NPC1L1, APOB, LDLR, APOC3, PCSK9	9
GO:0010742	macrophage derived foam cell differentiation	SOAT2, PPARG, SOAT1	14
GO:0051004	regulation of lipoprotein lipase activity	GPIHBP1, LMF1, ANGPTL3, PCSK6, ANGPTL8, LPL, FURIN, LIPC, PCSK5	13
GO:0065005	protein-lipid complex assembly	APOA4	22
GO:0097006	regulation of plasma lipoprotein particle levels	APOE, DGAT2	52
GO:0033993	response to lipid	SREBF1, INSIG2, SREBF2, CD36, ABCG1, PCK2, PCK1, PPARG, PPARD, PPARA	56

For each parent GO term of Table 3.13 and Table 3.14 there is a set of child GO terms, which are individually associated to several target genes, and that can be found in Annex 8.

Table 3.15. List of the other parent GO terms in BP functional domain, with the associated target genes. Ngc: number of genes associated with child terms of each parent term.

ID	GO term name	Genes	Ngc
GO:0007596	blood coagulation	HBB, P2RX4, P2RX2, PLAT, P2RX3, P2RX1, P2RX6, FGB, FGA, P2RX5, FGG, PLG	32
GO:0006811	ion transport	P2RX3, KCNMB2, P2RX1, ITPR1, ITPR2, P2RX2, ATP2B2, ATP2A1, TCN2, TRPC7, P2RX5, KCNMB3, TRPC6, ATP2A3, P2RX6, ORAI1, ATP2B4, TRPC3, ITPR3, ATP2B1, P2RX4, TCN1, SLC8A2, KCNMA1, SLC8A3, KCNMB4, ATP2B3, SLC8A1	27
GO:0044267	cellular protein metabolic process	FGG, SAA1, MMP1, PLG, FGA	63
GO:0006351	transcription, DNA-templated	MYC, MED20, MTF1, TEAD4	114
GO:0001666	response to hypoxia	ITPR2, SOD3, P2RX3, ITPR1, MTHFR, KCNMA1, PLAT, ARNT, P2RX2, SLC8A1	21
GO:0007165	signal transduction	TRAF4, P2RX6, ATF6, GNG3, ITPR2, GNB4, SRI, P2RX4, P2RX1, PPP2R5D, GNG2, PRKG2, GNG13, PRKG1, GNB2, GNGT2, GNAS, AMFR, P2RX5, ITPR1, PPP2R5C, GNG4, GNG11, PPP2R5E, P2RX3, GUCY1A2, CLOCK, PPP2R5A, GNG8, GNG10, GNGT1, GNG5, GNB1, GNB5, GSK3A, GNG7, PPP2R5B	57

Table 3.16. List of the other parent GO terms in MF functional domain, with the associated target genes. Ngc: number of genes associated with child terms of each parent term. **(Continued in the next page)**

ID	GO term name	Genes	Ngc
GO:0003677	DNA binding	TEAD1, ESRRA, ATF6, MED6, CLOCK, TEAD3, MYC, AHR, NRF1, TEAD4, TEAD2, THRAP3, ARNTL, SPI1, RXRB, CREB3L3, RXRG, MTF1, ARNT	43
GO:0003700	DNA-binding transcription factor activity	AHR, ARNTL, ESRRA, MYC, CLOCK, TEAD1, TEAD4, SPI1, ARNT, RXRB, TEAD2, ATF6, CREB3L3, NRF1, MTF1, TEAD3, RXRG	41
GO:0046872	metal ion binding	SLC8A1, SEC24B, SLC8A2, ATP2A3, PPP1CB, GPD2, MMP1, MTF1, TRAF4, SEC24D, CALM1, MCEE, PPP1CC, RXRB, RNF139, FGG, ATP2B1, ATP2B3, ATP2B2, PPP1CA, CALM3, GNAS, MTR, ESRRA, FGA, SIRT6, AMFR, CBS, ATP2B4, HBA1, SEC23B, MAT1A, SEC23A, CALM2, RXRG, SEC24C, TCN2, HBB, KCNMA1, SOD1, PPP2CA, SLC8A3, PPP2CB, SRI, ATP2A1, SAR1B, SOD3	56
GO:0000166	nucleotide binding	P2RX2, SAR1A, GUCY1A2, PIK3CA, P2RX3, PDPK1, INSR, PRKG1, CDK1, GNAS, ATP2A3, SAR1B, ATP2B4, PRKG2, GSK3A, ATP2B3, ACSS1, THRAP3, ATP2B1, P2RX5, MMAB, ILK, MAT1A, FGR, CDK8, ATP2A1, ATP2B2, P2RX4	70

ID	GO term name	Genes	Ngc
GO:0005515	protein binding	CIDEC, A2M, MAT1A, ITPR3, SAA4, ARNTL, CCND1, MDH1, ATF6, TRPC6, SOD1, MED7, GNB2, MED8, SIRT6, IFNG, MYC, FAM120B, ILK, GNAS, MED11, PLAT, MED12, YAP1, PLG, GNG4, INSR, PRKG1, SEC24C, SEC24D, MED14, MED16, FGA, ITPR1, RXRB, FGB, MED18, NRF1, MED19, GNG13, AHR, FGG, ARNT, FGR, MED23, MED22, MED26, ATP2B2, GNG8, P2RX1, CLOCK, RBP4, SRI, MED30, KCNMA1, MTR, MED31, PPP2R1B, MED4, TRPC3, MED6, PPP1CC, ATP2B1, PIK3CA, MED9, ATP2A1, SAA2, RNF139, GSK3A, SOD3, MMAB, CCNC, TEAD3, MED17, GUCY1A2, WWTR1, GNB1, SEC23B, CREB3L3, GNG3, PPP2R5D, HBB, PPP2R5E, PPP2CA, GNB4, SPI1, GNB5, ACSS1, AMFR, ATP2B3, GNG10, PPP2R1A, SDC1, GNG11, PPP1CB, LMF2, SHMT2, ESRRA, GNG2, PPP2R5C, RXRG, GNG5, ORAI2, GNG7, CALM3, MED27, CALM1, GNGT1, MED29, SAR1B, MCEE, ATP2B4, PDPK1, TEAD4, KCNMB4, KLK15, MTRR, TCN2, TEAD1, MTF1, HBA1, SEC13, TEAD2, PPP2CB, CDK1, PPP2R5A, AQP7, MED20, CALM2, TRAF4, MED10, NCOA3, SLC8A1, SEC23A, PPP2R5B, CTH, P2RX4, SEC31B, ORAI1, MED15, THRAP3, SAR1A, CBS, MED21, PPP1CA, SEC24B, CDK8, MED28, SEC31A, PTPN6, TRPC7	150
GO:0016740	transferase activity	SHMT2, NCOA3, RNF139, PRKG2, SIRT6, CDK8, INSR, ILK, PRKG1, AMFR, FGR, MMAB, PDPK1, CDK1, MAT1A, GSK3A, PIK3CA, MTR, CLOCK	50

For each parent GO term of Table 3.15 and Table 3.16 there is a set of child terms, which are individually associated to several target genes, and that can be found in Annex 9.

3.2. Identification of candidate genes for future genetic studies

For identification of the target genes with the best potential to be useful in future GWAS and other molecular studies of interest for dyslipidaemias, we considered the phenotypic and functional information collected for each gene. Accordingly, a total of 41 genes associated with at least one of the nine selected GWAS traits and showing relation with lipid-specific GO terms of both GO domains were identified as candidate genes (Table 3.17). These genes might be particularly important for a better understanding of the polygenic contribution to dyslipidaemia and help unrevealing differences in phenotype, mainly in biochemical profile, between patients with a similar diagnosis.

Table 3.17. Candidate genes with Ensembl ID, official gene symbol and full name.

ID	Gene	Full name
ENSG00000165029	ABCA1	ATP binding cassette subfamily A member 1
ENSG00000005471	ABCB4	ATP binding cassette subfamily B member 4
ENSG00000160179	ABCG1	ATP binding cassette subfamily G member 1
ENSG00000138075	ABCG5	ATP binding cassette subfamily G member 5
ENSG00000143921	ABCG8	ATP binding cassette subfamily G member 8
ENSG00000163631	ALB	albumin
ENSG00000118137	APOA1	apolipoprotein A1
ENSG00000110244	APOA4	apolipoprotein A4
ENSG00000110243	APOA5	apolipoprotein A5
ENSG00000084674	APOB	apolipoprotein B
ENSG00000130208	APOC1	apolipoprotein C1
ENSG00000234906	APOC2	apolipoprotein C2
ENSG00000110245	APOC3	apolipoprotein C3
ENSG00000267467	APOC4	apolipoprotein C4
ENSG00000130203	APOE	apolipoprotein E
ENSG00000135218	CD36	CD36 molecule
ENSG00000087237	CETP	cholesteryl ester transfer protein
ENSG00000277494	GPIHBP1	glycosylphosphatidylinositol anchored high density lipoprotein binding protein 1
ENSG00000115677	HDLBP	high density lipoprotein binding protein
ENSG00000125629	INSIG2	insulin induced gene 2
ENSG00000130164	LDLR	low density lipoprotein receptor
ENSG00000157978	LDLRAP1	low density lipoprotein receptor adaptor protein 1
ENSG00000107798	LIPA	lipase A, lysosomal acid type
ENSG00000166035	LIPC	lipase C, hepatic type
ENSG00000182333	LIPF	lipase F, gastric type
ENSG00000101670	LIPG	lipase G, endothelial type
ENSG00000175445	LPL	lipoprotein lipase
ENSG00000123384	LRP1	LDL receptor related protein 1
ENSG00000138823	MTTP	microsomal triglyceride transfer protein
ENSG00000144959	NCEH1	neutral cholesteryl ester hydrolase 1
ENSG00000141458	NPC1	NPC intracellular cholesterol transporter 1
ENSG00000015520	NPC1L1	NPC1 like intracellular cholesterol transporter 1
ENSG00000025434	NR1H3	nuclear receptor subfamily 1 group H member 3
ENSG00000169174	PCSK9	proprotein convertase subtilisin/kexin type 9
ENSG00000166819	PLIN1	perilipin 1
ENSG00000100979	PLTP	phospholipid transfer protein
ENSG00000266200	PNLIPRP2	pancreatic lipase related protein 2 gene/pseudogene
ENSG00000186951	PPARA	peroxisome proliferator activated receptor alpha
ENSG00000069667	RORA	RAR related orphan receptor A
ENSG00000073060	SCARB1	scavenger receptor class B member 1
ENSG00000147852	VLDLR	very low density lipoprotein receptor

Among candidate genes (Table 3.17), there are some well-known players in lipid metabolism and dyslipidaemia, including the FH-related genes (*LDLR*, *APOB*, *PCSK9*) [9], and those associated to FH phenocopies (monogenic dyslipidaemias with a similar phenotype to FH but resulting from variants in other lipid-related genes) such as *LDLRAP1*, *LIPA*, *ABCG5* and *ABCG8* genes [33]. Other genes not commonly associated with dyslipidaemia and present in this list (Table 3.17) are, for example, *HDLBP*, *PLIN1*, *RORA*, or *INSIG2*. Still, most of the candidate genes are related to lipoproteins receptors and ligands, apolipoproteins, and lipases.

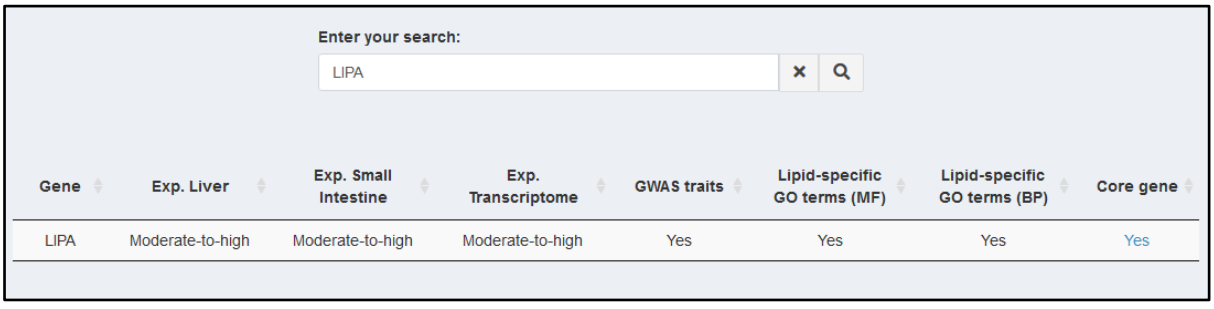
3.3. Development of a shiny app for hosting the new lipid knowledge base

Considering the potential utility for the scientific community of the new lipid knowledge base, this was hosted by a new freely available website that was developed using the R shiny application (see methods). This shiny app will allow the user to visualise and interact with the data present in the knowledge base, and it can be found using the following link: https://shiny.campus.ciencias.ulisboa.pt/rnasysbio/MylipidgenesKB/. The main aim of this knowledge base is to offer access to a universe of target genes involved in central pathways of lipid and lipoprotein metabolism, and that may contribute to unravel the biological context of dyslipidaemia and improve the integration of lipid knowledge. This new website allows the user to directly access other public available databases that contained additional information regarding any of these genes, including pathways databases (Wikipathways and Reactome), GeneCards for diverse information, HUGO gene nomenclature committee (HGNC) database for checking multiple ID systems, and PubMed for allowing to find the most important literature and recently published articles. Considering the interaction offered by this shiny app, the user can visualise gene expression profiles among target genes (maximum of 20 genes at the same time for comparison); download lipid specific lists of GO terms for BP and MF domains but also other non-lipid associated GO terms that are representative of target genes, with GO terms grouped according to their gene ontology hierarchical relations and accessible links for QuickGO database; visualise gene interactions among networks of target genes that can be selected according to associated GWAS traits. In addition, there is the possibility to search for any gene in the homepage (Figure 3.15) to check the information available for that gene in the knowledge base. The following examples help to understand the utility of this knowledge base as an online resource for the scientific community.

Figure 3.15. Homepage of MylipidgenesKB showing a search toolbar for genes and a lateral menu (left panel).

3.3.1. Example A: Retrieve metabolic, transcriptomic, phenotypic, and functional information for *LIPA* gene

In this example, the aim is to get an overview of the gene *LIPA* regarding metabolic pathways where it is involved, gene expression level in lipid-related tissues, associated GWAS traits and GO terms, considering the biological context of the knowledge base. Thus, the user will be able to look for potential gene expression, phenotypic/disease, and functional patterns, while comparing *LIPA* with other lipid-related genes. For this, the first step is to search for *LIPA* in the homepage and get a quick summary of the available information for this gene. As shown in Figure 3.16, *LIPA* has a moderate-to-high expression in both tissues of interest and transcriptome, it presents associations to selected GWAS traits and to lipid-specific GO terms of both GO domains, which makes *LIPA* a core gene. The designation "core" was only used in the shiny app "MylipidgenesKB" as an alternative name for candidate genes.

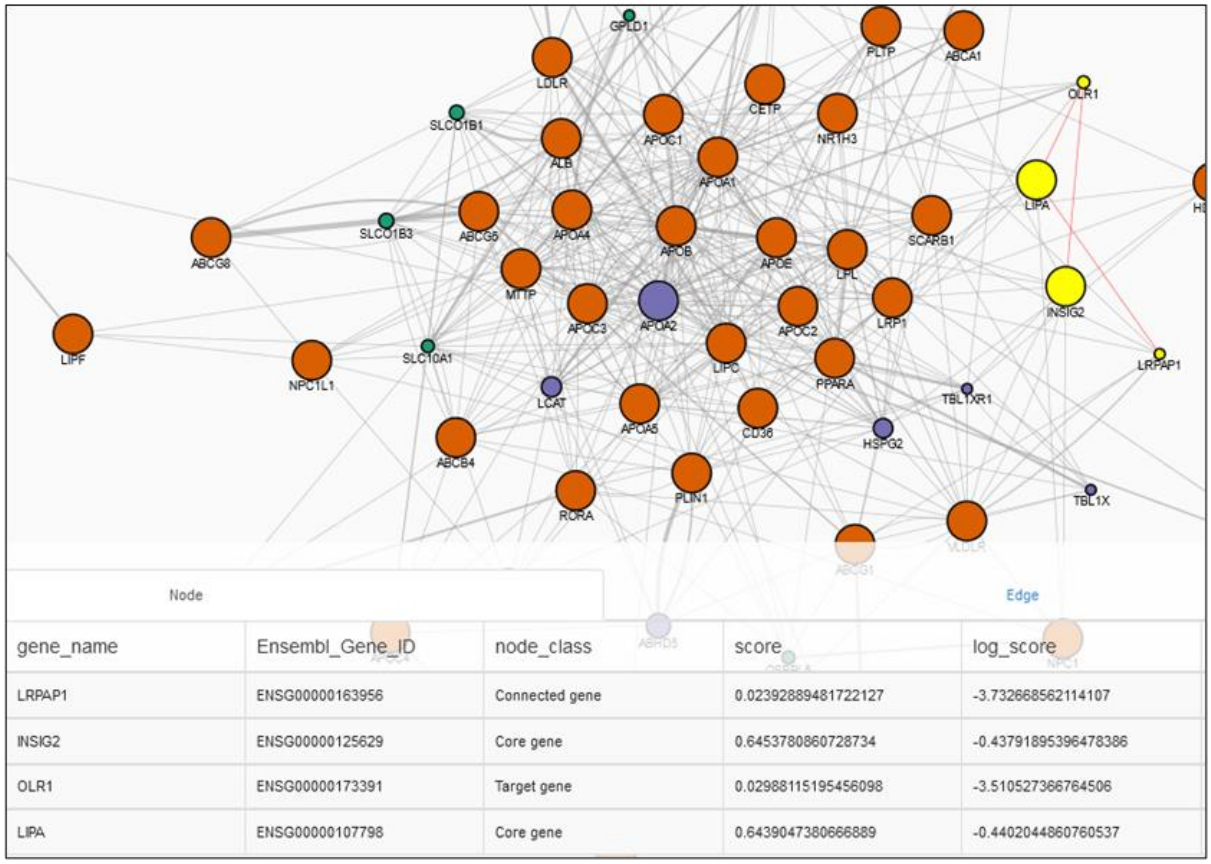


Gene	Exp. Liver	Exp. Small Intestine	Exp. Transcriptome	GWAS traits	Lipid-specific GO terms (MF)	Lipid-specific GO terms (BP)	Core gene
LIPA	Moderate-to-high	Moderate-to-high	Moderate-to-high	Yes	Yes	Yes	Yes

Figure 3.16. Search output for *LIPA* in MylipidgenesKB homepage showing a summary table with the available information in this knowledge base.

By opening the link available in the column "Core gene" (Figure 3.16), the user may access a gene interactions network of all core genes, including *LIPA*. This network also comprises other target genes (not core) and genes absent from target list (connected genes) that have associations with any of the core genes. As shown in Figure 3.17, the user can manually select any gene in the network, for example *LIPA*, and access a table showing additional information regarding

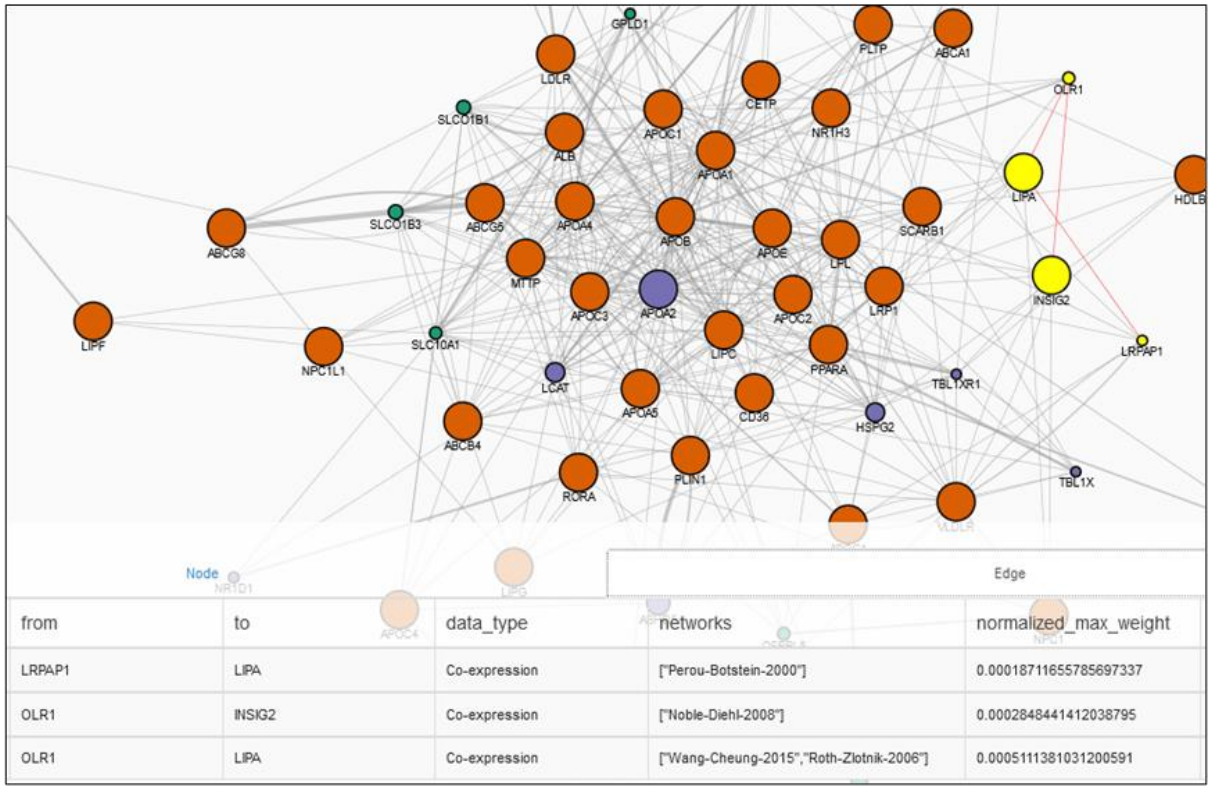
the selected gene (table panel “node”). When more than one gene is selected (Figure 3.18), this table also shows information regarding the interactions among selected genes (table panel “edge”). The information present in both table panels can be downloaded, considering selected genes and interactions.



gene_name	Ensembl_Gene_ID	node_class	score	log_score
LRPAP1	ENSG00000163956	Connected gene	0.02392889481722127	-3.732668562114107
INSIG2	ENSG00000125629	Core gene	0.6453780860728734	-0.43791895396478386
OLR1	ENSG00000173391	Target gene	0.02988115195456098	-3.510527366764506
LIPA	ENSG00000107798	Core gene	0.6439047380666889	-0.4402044860760537

Figure 3.17. Core genes network with manually selected genes in yellow, including *LIPA*, and selected gene interactions in red. For selected genes, additional information is shown in the table panel “node” in the lower side of the figure. Score is higher as the number of interactions with core genes increases, which corresponds to a higher node size. The network is not fully shown in this figure. Unselected genes are present in different colours as follows: core genes (orange), target genes (green), connected genes (purple).

Considering Figure 3.17, two out of the four selected genes are core genes, *LIPA* and *INSIG2*, which present a higher score than *LRPAP1* (target gene) and *OLR1* (connected gene). This means that the two core genes have a higher number of interactions in this network in comparison to the other selected genes.



from	to	data_type	networks	normalized_max_weight
LRPAP1	LIPA	Co-expression	["Perou-Botstein-2000"]	0.00018711655785697337
OLR1	INSIG2	Co-expression	["Noble-Diehl-2008"]	0.0002848441412038795
OLR1	LIPA	Co-expression	["Wang-Cheung-2015","Roth-Zlotnik-2006"]	0.0005111381031200591

Figure 3.18. Core genes network showing table panel "edge" in the lower side of the figure. In the table, "data_type" corresponds to the nature of gene interaction (i.e., co-expression, physical interactions, pathway, genetic interactions), and "networks" presents published networks that GeneMANIA uses for establishing the gene interactions of the core network. The value of "normalized_max_weight" represents the strength of each interaction, with a higher weight corresponding to a stronger interaction between genes and a higher edge width. Selected genes in yellow, including *LIPA*, and selected gene interactions in red. The network is not fully shown in this figure. Unselected genes are present in different colours as follows: core genes (orange), target genes (green), connected genes (purple).

As present in Figure 3.18, all the selected gene interactions were based on co-expression profiles and the strongest of the three is between *OLR1* and *LIPA*, followed by *OLR1 – INSIG2* and *LRPAP1 – LIPA* interactions.

Then, by searching for "LIPA" in the target gene list, the user can access gene full name, associated metabolic pathways (with link to Wikipathways or Reactome), and external links for other databases that may offer additional information (Figure 3.19). The full target genes list can be downloaded as a .csv file.

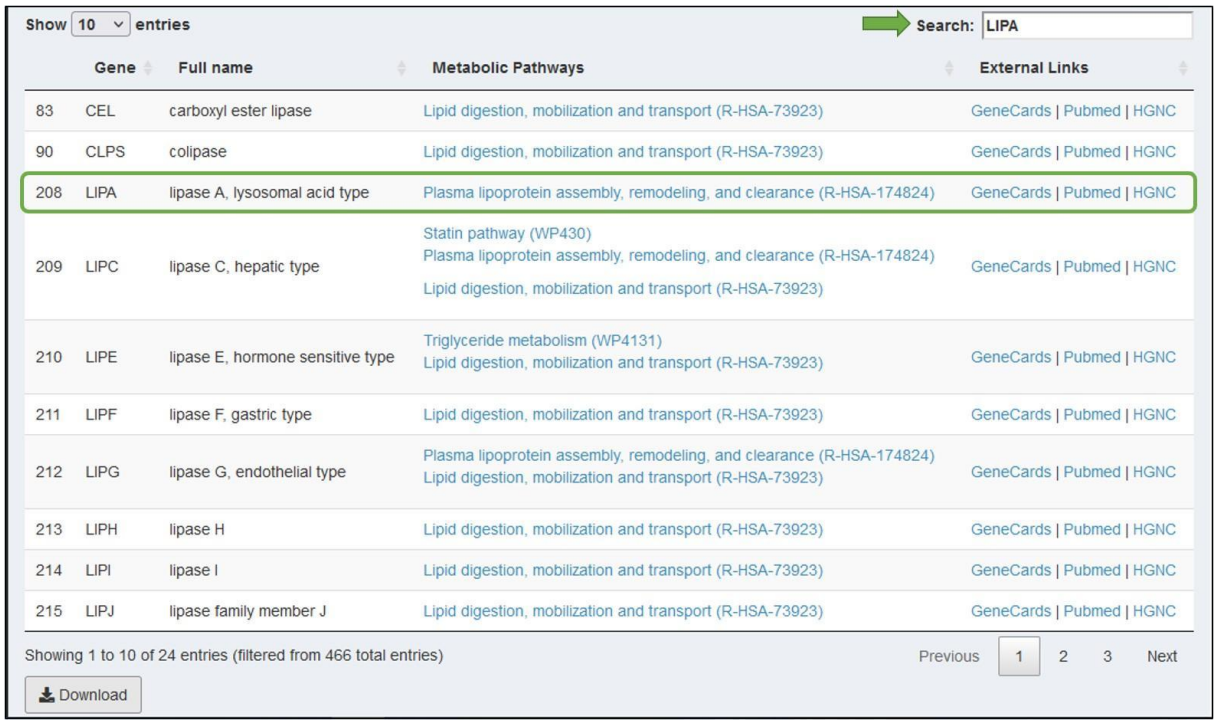


Figure 3.19. Target genes list filtered by searching "LIPA" in the whole table. The correspondent row to the *LIPA* gene is in green contour.

MylipidgenesKB allows the user to compare gene expression profiles of a maximum of 20 genes at the same time, for tissues of interest and transcriptome. In this example, *LIPA* expression is compared with two genes previously considered in the analysis of the core genes network – *OLR1* and *INSIG2* (Figure 3.20), which are also part of the target gene list.

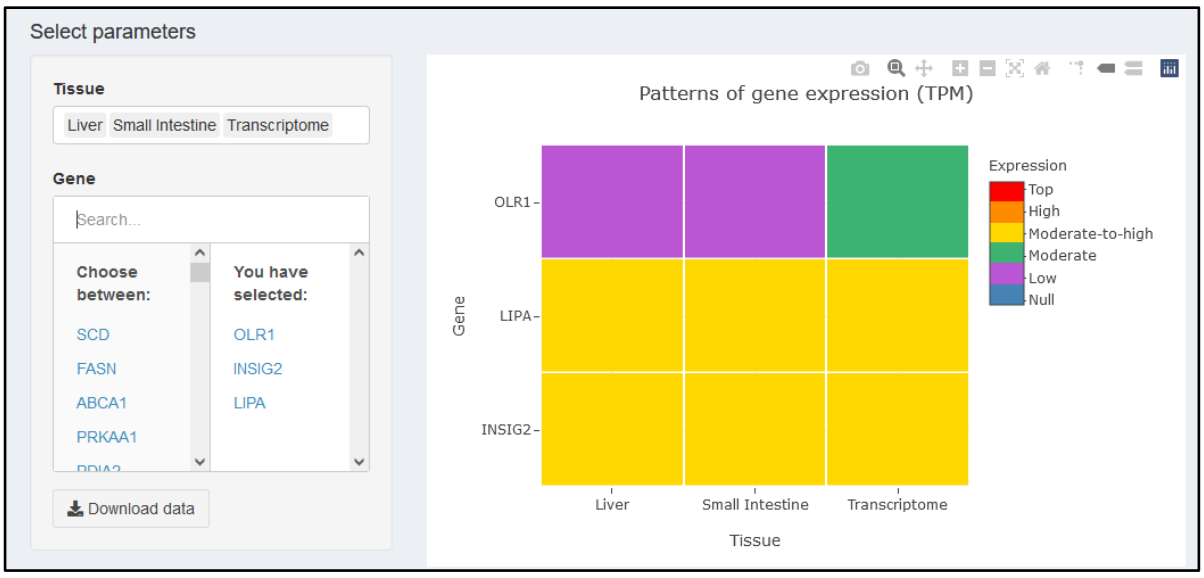


Figure 3.20. Comparison of gene expression pattern of *LIPA* and two interacting genes in the core network – *OLR1* and *INSIG2*. In the left panel, the user can select the tissue (among liver, small intestine, and transcriptome) and genes from the target list. The selected data can be downloaded as a .csv file. In the right panel, a heatmap shows the pattern of gene expression across the selected tissues and genes, where different colours correspond to different expression categories. The icons present above the heatmap correspond to the following from left to right: download plot as .png file, zoom, pan, zoom in, zoom out, autoscale, reset axes, toggle spike lines, show closest data on hover (selected), compare data on hover, and produced with plotly.

As shown in Figure 3.20, *LIPA* and *INSIG2* have a similar gene expression pattern in tissues of interest, and this is also similar to transcriptome estimated expression. Conversely, *OLR1* presents an expression level in tissues of interest lower than in transcriptome, and even lower in comparison to *LIPA* and *INSIG2* levels.

In the "GWAS" section, the user can start by checking the network for all trait frequencies in the subsection "by trait number", to find how many and which GWAS traits are associated with *LIPA* (Figure 3.21).

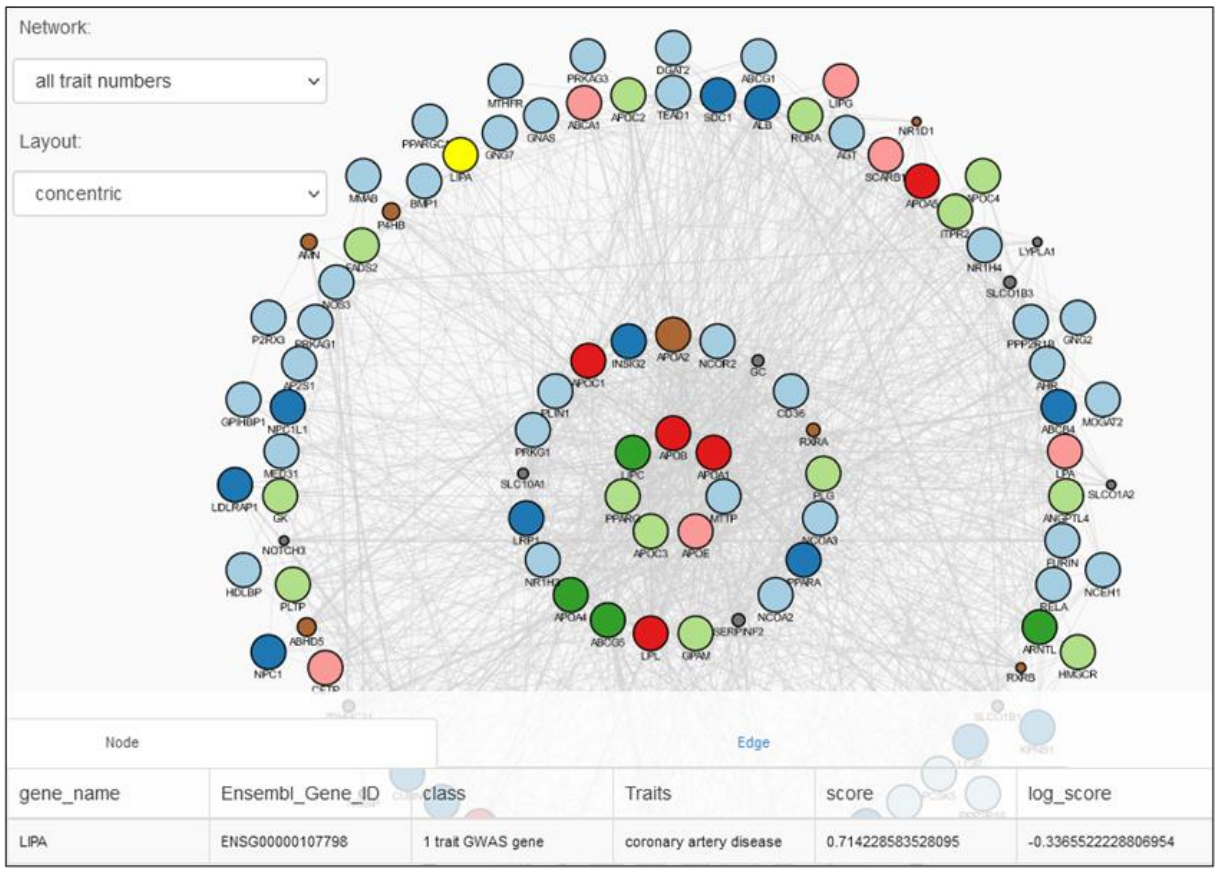


Figure 3.21. Gene interactions network comprising genes with associated GWAS traits (GWAS genes) and other interacting genes (target and connected genes), and having *LIPA* manually selected (yellow). GWAS genes are coloured according to the number of associated traits as follows: 1 trait (light blue), 2 traits (dark blue), 3 traits (light green), 4 traits (dark green), 5 traits (light red), 6 traits (dark red). Target and connected genes are shown in brown and grey, respectively. In this figure, the selected network comprises all GWAS genes independently of the number of associated traits. The table panel "node" presents additional information for *LIPA* (downside of figure).

According to Figure 3.21, *LIPA* is associated with one GWAS trait – coronary artery disease. Next, the user can check the network of "1 trait GWAS genes" to observe the connections between *LIPA* and other genes with only one associated trait, looking which trait is associated to these genes and comparing score values (table panel "node") and strength of gene interactions (table panel "edge"). In addition, at the subsection "by trait", the user can look at the network

of all genes associated with coronary artery disease that includes *LIPA*, and check interactions between genes that are associated with the same trait.

Considering functional information, *LIPA* can be searched within parent GO terms tables for both GO domains available in MylipidgenesKB – BP (Figure 3.22) and MF (Figure 3.23). *LIPA* is only present in lipid-specific tables and no "other GO terms" are available for this gene.

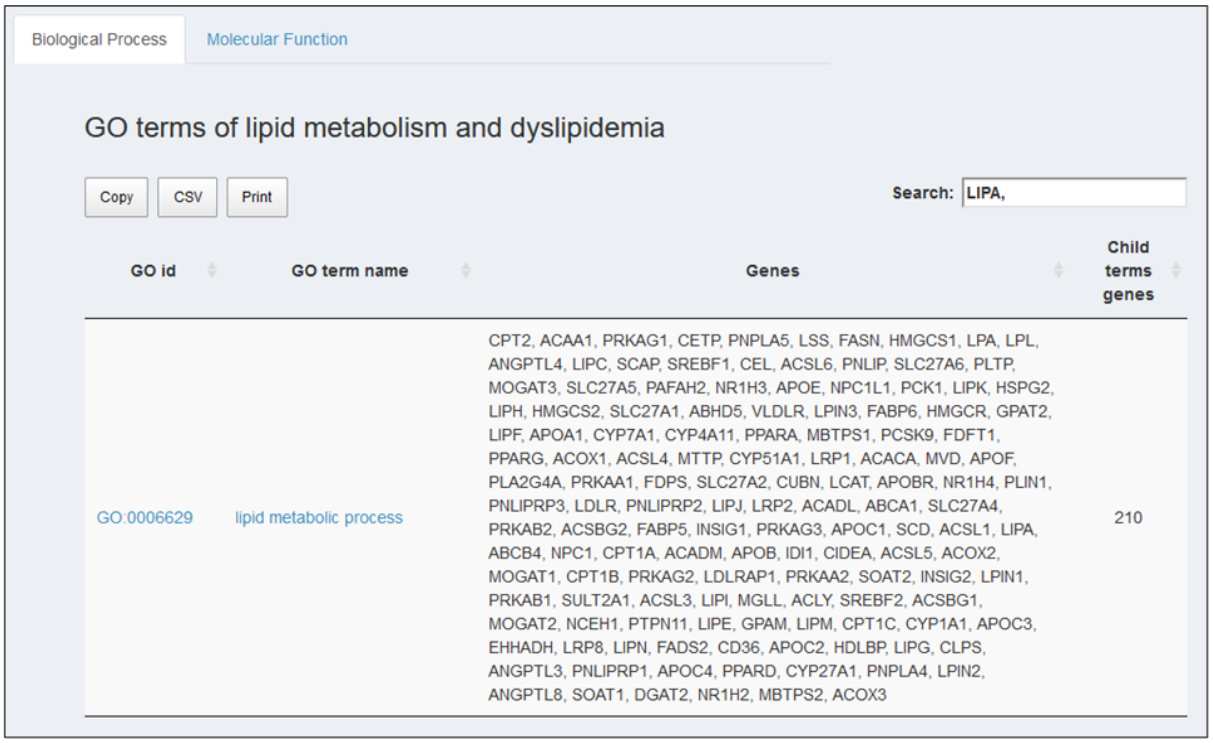


Figure 3.22. Lipid-specific GO terms associated with *LIPA* for BP domain. Adding a comma after "LIPA" in the search field avoids a mismatch between "LIPA" and "lipase". GO id opens a link for QuickGO database, GO term name opens a separate window with a table of child terms. Search output table can be copied, downloaded as a .csv file or printed. The column "Child terms genes" corresponds to the number of genes associated with all child terms of each parent term.

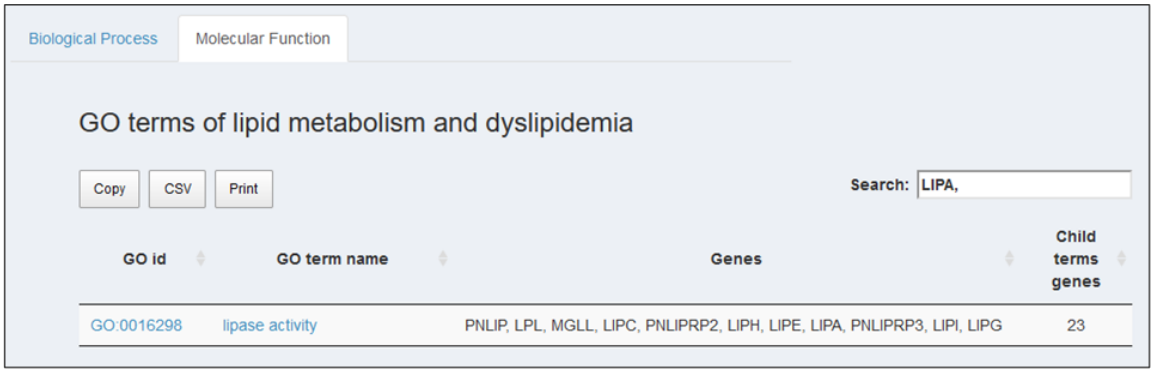
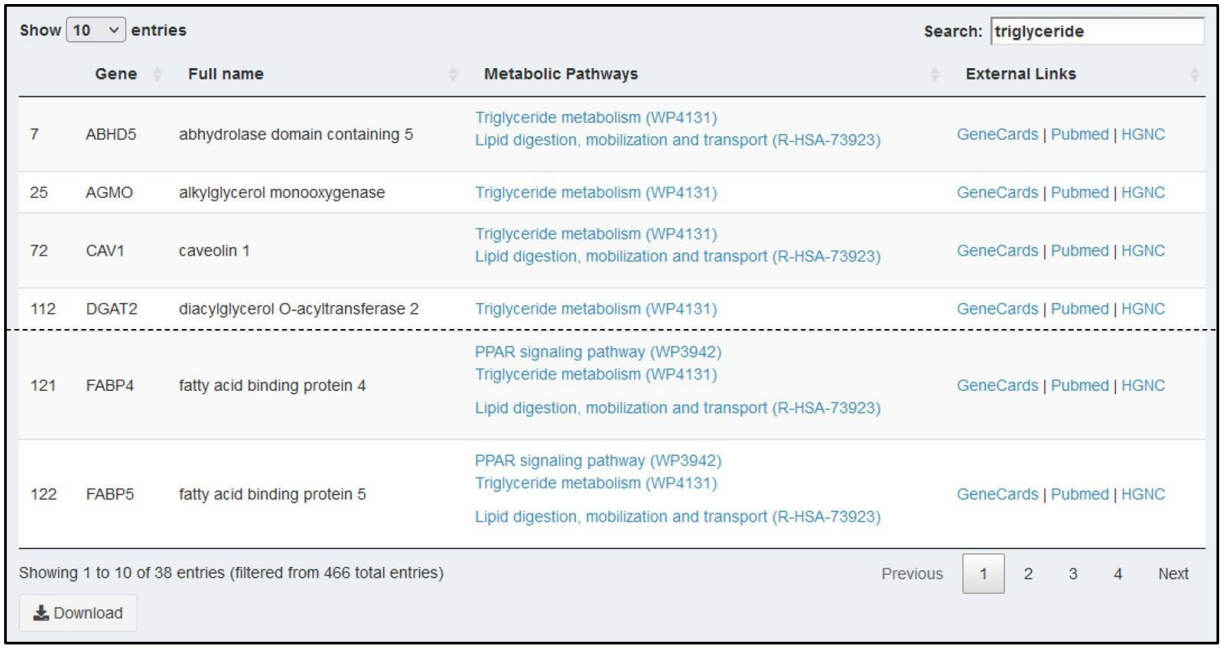


Figure 3.23. Lipid-specific GO terms associated with *LIPA* for MF domain. Adding a comma after "LIPA" in the search field avoids a mismatch between "LIPA" and "lipase". GO id opens a link for QuickGO database, GO term name opens a separate window with a table of child terms. Search output table can be copied, downloaded as a .csv file or printed. The column "Child terms genes" corresponds to the number of genes associated with all child terms of each parent term.

For each GO domain, there is a parent GO term associated with *LIPA*, namely "lipid metabolic process" (BP) and "lipase activity" (MF).

3.3.2. Example B: Retrieve target genes related to a lipid parameter

In this example, the user aims to identify the target genes related to TG metabolism and explore potential patterns of gene expression, associated GO terms and phenotypic/disease traits of these genes. TG are a commonly measured lipid parameter with a central role in lipid metabolism and dyslipidaemia biological background. Given the fact that the search field in the homepage only accepts gene symbols, the user should start by searching "triglyceride" in the gene list table and check which target genes are related to triglycerides metabolism (Figure 3.24).


Show 10 entries Search: triglyceride

	Gene	Full name	Metabolic Pathways	External Links
7	ABHD5	abhydrolase domain containing 5	Triglyceride metabolism (WP4131) Lipid digestion, mobilization and transport (R-HSA-73923)	GeneCards | Pubmed | HGNC
25	AGMO	alkylglycerol monooxygenase	Triglyceride metabolism (WP4131)	GeneCards | Pubmed | HGNC
72	CAV1	caveolin 1	Triglyceride metabolism (WP4131) Lipid digestion, mobilization and transport (R-HSA-73923)	GeneCards | Pubmed | HGNC
112	DGAT2	diacylglycerol O-acyltransferase 2	Triglyceride metabolism (WP4131)	GeneCards | Pubmed | HGNC
121	FABP4	fatty acid binding protein 4	PPAR signaling pathway (WP3942) Triglyceride metabolism (WP4131) Lipid digestion, mobilization and transport (R-HSA-73923)	GeneCards | Pubmed | HGNC
122	FABP5	fatty acid binding protein 5	PPAR signaling pathway (WP3942) Triglyceride metabolism (WP4131) Lipid digestion, mobilization and transport (R-HSA-73923)	GeneCards | Pubmed | HGNC

Showing 1 to 10 of 38 entries (filtered from 466 total entries) Previous 1 2 3 4 Next

Download

Figure 3.24. Target genes list filtered by searching "triglyceride" in the whole table. The pointed line shows that only some of the rows are represented.

Considering the target genes found to be involved in triglycerides metabolism, the user can search these genes at homepage to get a quick summary of the available information and compare gene expression patterns in the "tissue expression" section. In addition, following what was done for example A, TG related genes can be found in the network comprising all GWAS genes (Figure 3.21). In the other subsection ("by trait"), the user can select the GWAS network of hypertriglyceridaemia (Figure 3.25), which is composed of genes associated with high levels of TG and their interacting genes.

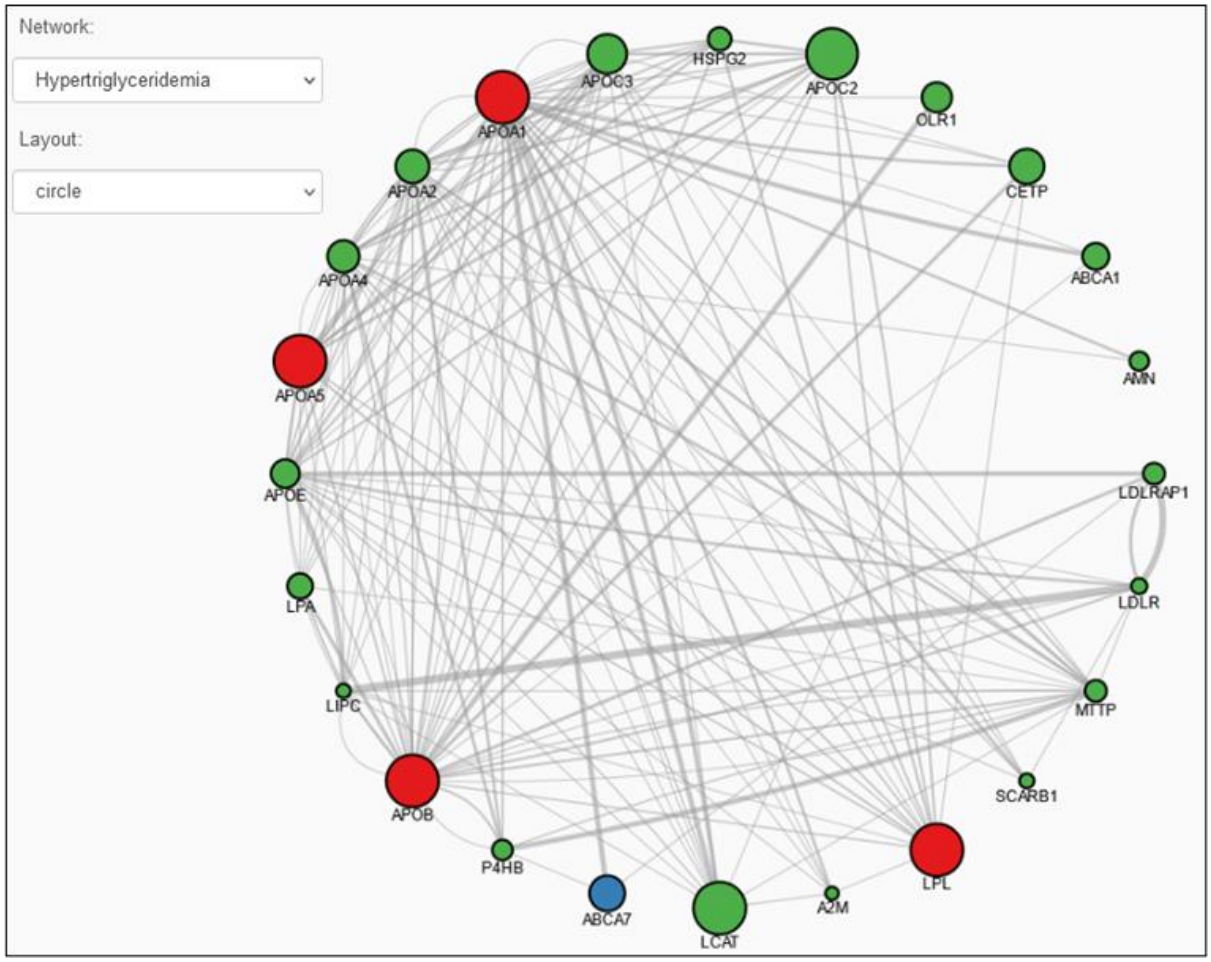


Figure 3.25. Gene interactions network for hypertriglyceridaemia. In the network menu (upper left corner), users can choose among the networks related to each GWAS trait. Layout menu (upper left corner) offers different options for network visualisation. GWAS genes are present in red, target genes in green, and connected genes in blue.

In all tables of the section "Gene Ontology", the user can search for "triglyceride" or TG related genes previously identified in the target gene list. In some cases, no search results are found in parent GO terms tables, but it does not mean that there are no child GO terms associated with TG levels or TG related genes. Each parent term has a link in the term name that opens a new window with a table that includes all child terms selected for that parent term. As the tables of parent terms, the tables with child terms allow users to search all their content and download, print and/or copy the terms that result from that search, and also present links in the "GO id" column for QuickGO database.

The new lipid knowledge base developed in this project is the result of an integrative analysis of diverse publicly available data, including gene expression, associated phenotypic/disease traits and functional terms, based on the exploration of lipid metabolic pathways that were selected considering previous knowledge and the results of machine-learning analysis applied on a known dyslipidaemic dataset. The definition of a universe of lipid-related genes, taking

into account the selected metabolic pathways, allowed the establishment of molecular interactions relevant for the dyslipidaemic context and future genetic studies. Therefore, this knowledge base may be useful to identify additional biomarkers and genes of interest, for a better discrimination between dyslipidaemic patients, while contributing to the improvement of lipid knowledge integration.

Chapter 4

Discussion and final remarks

1. Training of classification models that improve distinction between FH+ and FH- individuals

Given the high risk for severe CVD at an early age and the benefits of early therapeutic intervention, the identification of children carrying monogenic FH variants is of extreme importance. Biochemical identification of dyslipidaemic subjects in clinical practice usually relies on the analysis of serum levels for total cholesterol, HDL-C, TG, LDL-C and eventually apoA-I and apoB [51], [62]. Although these biochemical markers allow for a relatively sensitive screening of individuals at risk for CVD, including FH candidates, their specificity in distinguishing monogenic individuals is very low [145]. In addition, recent studies show that many children do not comply with multiple parameters of clinical diagnostic criteria, including the presence of family history of hypercholesterolaemia/CVD or LDL-C levels above the defined cut-offs [51], [61]. Screening for genetic variants was therefore recommended as standard of care for patients with definite or probable FH by an international Expert Consensus Panel [61]. However, the diagnostic yield of these screening programs is low [146], ranging between 20% to 80% [147], as a high number of suspected patients suffer from a polygenic condition [61]. Thus, the development of robust approaches that can contribute to increase this yield is critical to support a widespread use of FH genetic testing, with a considerable reduction of the resulting burden on health systems.

In this thesis, ML-based methods were applied to perform a thorough analysis of the extended lipid profiles of the PFHS-ped dataset. This approach was adopted taking into account the hypothesis that using an extended lipid profile would confer an additional layer of information, supporting a more accurate identification of FH+ subjects, leading to the identification of novel clinically relevant biomarkers. Multiple “training” sets comprising different combinations of biochemical parameters were used to train classification models to distinguish FH+ and FH- individuals, followed by an assessment of performance on independent “testing” sets. For comparison purposes, similar models using only TC and LDL-C were trained. Predictions of FH+ and FH- status for the same group of patients were performed using the two best models, SB models and standard SB criteria cut-offs (Table 3.4). Results show that modelling can considerably improve the specific identification of FH+ individuals and the PPV, with a limited impact on the high sensitivity afforded by SB cut-off criteria. Furthermore, the inclusion of extended lipid parameters contributes to an improved patient identification.

The best ranking model Imp_B uses apoB/apoA-I and TG/apoB ratios, in addition to LDL-C levels, to generate predictions with the highest sensitivity values. Of note, LDL-C levels used in this study were directly determined and thus their accuracy is not affected by TG levels. The

current guidelines for dyslipidaemia already recommend the determination of LDL-C, TG and apoB in all dyslipidaemic individuals [62]. Like the TC/HDL-C, the apoB/apoA-I ratio has been linked to cardiovascular risk [148]. Indeed, a previous study identified the apoB/apoA-I ratio as a potential biomarker for FH [35]. The TG/apoB ratio was selected both in the first and second ranked models, the later delivering the highest specificity and PPV. This model further includes two "Basic" biochemical parameters (TC and TC/HDL-C) and LDL1 from Lipoprint analysis (see methods). Of note, LDL1 is the most commonly selected biochemical parameter across all top 10 models, suggesting it holds relevant information for the specific identification of FH+ individuals.

All the selected parameters make sense in the frame of the current understanding of lipid metabolism and the biology behind hypercholesterolaemia, while providing new insights and hypotheses into underlying metabolic differences in FH+ and FH- dyslipidaemic cases. Thus, whereas the altered blood lipid parameters of FH+ individuals can be explained by the well-established impairment of LDL internalisation in the liver, our results suggest that dyslipidaemia in FH- individuals predominantly involves an imbalance of TG-related pathways. This can either be due to environmental causes like a high dietary TG intake, or to a polygenic background affecting a distinct group of genes. Although the polygenic basis of hypercholesterolaemia in FH- individuals has been widely suggested [35], [62], [67], [149], most studies focus on the involvement of the same pathways affected in FH+ individuals [36], [68], [150]. Our results further suggest that altered HL-dependent lipolysis of IDL and LDL particles is a relevant phenomenon in FH, leading to a change in the relative proportions of the different LDL subfractions, which in turn may have an impact on disease severity. Indeed, several studies have associated altered HL activity or expression, namely in association to genetic polymorphisms, to more severe FH phenotypes [151], [152].

Overall, the obtained results suggest that modelling, together with the inclusion of novel lipid parameters, can support an improved classification of FH+ and FH- individuals, with a significant impact on the yield of genetic screening programs and corresponding costs. The top models can already be used by clinicians to obtain more precise estimates of the likelihood that their patients are FH+ in comparison to SB criteria. The PPVs and NPVs described in Table 3.4 should be taken into consideration when interpreting results. All the required information for their application is provided in GitHub (see link in results). The availability of larger patient datasets will be crucial to identify which of the new, non-standard parameters used by the trained models will be worth incorporating into clinical practice, as well as for investigating if the proposed metabolic differences are observed in other populations.

2. Identification of different dyslipidaemic profiles among individuals by a hierarchical clustering analysis

Similar to the results obtained with the training of classification models through the application of a ML-based approach, the performed clustering analysis revealed the considerable potential of novel methodologies, namely associated to data science, towards the identification of new biomarkers for clinical purposes. Indeed, the applied clustering approach was able to identify different lipid profiles among individuals beyond what was provided by the FH+/FH- classification. For every subset, clustering analysis revealed the presence of a third group of patients, with each of the three clusters presenting a distinct pattern regarding the prevalence of FH+ and FH- individuals. Then, it was possible to verify that in every subset, in addition to a cluster mainly constituted by FH+ patients and another cluster mostly composed by FH- individuals, there was a third cluster with a considerable number of both FH+ and FH- individuals, corresponding to a mixed population. Still, the differences among clusters were better defined in the "All" subset, thus emphasising the importance of using an extended lipid profile to improve the distinction between individuals.

Concerning cluster description through the analysis of the quantitative variables that most contribute for each cluster profile and for the cluster partition in three distinct groups, this allowed the identification of pathways of interest in the context of lipid metabolism. Hence, evidence suggests an important role of LDL/apoB pathway and TG metabolism as main contributors for the lipid profiles of the predominant FH+ and the predominant FH- clusters, respectively. The association of these pathways to each group of individuals (FH+ and FH-) was already established in a previous study analysing PFHS dataset [35]. The potential contribution of perturbed TG metabolism to the dyslipidaemic profile of FH- individuals was previously discussed in the section of modelling analysis on chapter 3 and 4. Of note, the considerable presence of parameters associated to LDL subfractions (result of the VLDL – IDL – LDL delipidation cascade) among the variables that best described the predominant FH+ cluster, which highlights once again the importance of an extended lipid profile to acquire a better distinction of individuals.

The description of clusters by categorical variables gave us additional information, allowing the identification of molecular, biochemical, and anthropomorphic patterns among clusters, including suggestions for a better understanding of the distribution of individuals within the mixed cluster, where no clear pattern was detected so far. Accordingly, the pattern identified in the predominant FH- cluster was associated with a significant presence of girls suffering from obesity with lower TC and LDL-C levels, in comparison to FH+ patients, with no fulfilment of

the cut-offs from SB criteria. In contrast, individuals of the predominant FH+ cluster presented higher levels of TC and LDL-C, with fulfilment of SB cut-offs, and *LDLR* pathogenic variants showing up to 20% of molecular activity in the affected allele. The polygenic LDL-C score was higher once we moved in the direction of the predominant FH- cluster. These evidences suggest a considerable contribution of polygenic and lifestyle factors for the development of dyslipidaemia in FH- individuals, as hypothesised by other studies [35], [36], [67]. The use of BMI class as supplementary variable emphasises the information given by BMI, which was one of the quantitative variables that presented a significant contribution to cluster partition. Given the fact that BMI class was based on WHO percentiles dependent of age and sex, it takes into consideration the different phases of development during childhood and adolescence. Regarding gender, we should take into account that despite its association with the predominant FH- cluster, this variable did not present a significant association with the cluster partition. Future studies using a bigger sample than PFHS-ped and enrolling also normolipidaemic individuals, aiming for a better representation of the general population, would be essential to validate gender as a discriminant factor of lipid levels. In relation to the mixed cluster, these individuals appeared to constitute a mixed phenotype, since their profile is milder than those from the predominant FH+ cluster but more severe than those from the predominant FH- cluster. Indeed, taking the dendrogram as reference, the degree of dyslipidaemia severity (mainly regarding levels of TC and LDL-C and results of the molecular study) appeared to be gradually increased from right to left, while the polygenic contribution appeared to be gradually increased from left to right.

Although Lipoprint assay allow us to acquire useful information regarding LDL subfractions, we should remember that it is a semiquantitative method and that the obtained measurements may not be as accurate as those achieved using quantitative methods (e.g., photometric test used for measuring parameters of "Advanced" profile) [16]. Indeed, FH+ patients present higher sdLDL concentrations when using a "daytona" assay (RX daytona+® analyser), in comparison to FH- individuals, which is expected regarding the role of sdLDL as pro-atherogenic particles within the context of FH. Conversely, Lipoprint results regarding the predominance of profile A among FH+ population and profile B within the predominant FH- cluster can be explained by an altered HL activity in FH+ patients (as already mentioned in this chapter) that results in higher prevalence of LDL1/LDL2 subfractions, and/or higher influence of lifestyle (e.g., TG-rich diets) and polygenic factors in FH- individuals [142], [149], [152], [153]. Of note, although FH- subjects present lower TC and LDL-C levels in comparison to FH+ patients, they are not normolipidaemic and usually tend to present borderline values of TC and LDL-C when they

fail to fulfil the cut-offs of SB criteria. This can be explained by a likely presence of a polygenic form of dyslipidaemia, resulting from the cumulative burden of several SNPs able to raise LDL-C levels [36], [67].

Further, the identification of the most specific and representative individuals of each cluster helped to reveal different dyslipidaemic profiles beyond the FH+/FH- classification. The predominant FH+ cluster was mainly characterised by FH+ individuals with TC and LDL-C levels above the SB cut-offs, *LDLR* pathogenic variants associated to a degree of molecular activity less or equal to 20% in the affected allele, moderate polygenic score and normal BMI. In contrast, the predominant FH- cluster was mainly characterised by FH- individuals with TC and LDL-C levels above the SB cut-offs, high polygenic score and obesity. Regarding the mixed cluster, the pattern found suggests a mixed phenotype with characteristics of both FH+ and FH- profiles. Despite the low number of individuals involved in this analysis (i.e., five individuals that best characterise each cluster), these findings are in agreement with the results previously discussed in this section.

The characterization of clusters regarding the contribution of each dimension used for HCPC analysis highlighted the influence of different lipid patterns in individual distribution. PC1 and PC2, considered the most informative components, were associated with the previous findings regarding the connection of LDL/apoB pathway and TG metabolism with predominant FH+ and FH- clusters, respectively. Although with a smaller level of contribution, PC4 revealed a strong association with the predominant FH- cluster, while their main contributors were parameters related to LDL subfractions and others like BMI and TG/apoB. The association between this cluster and delipidation cascade might be explained by the predominance of a Lipoprint profile B in FH- individuals, as formerly discussed in this section. In addition, the association of these individuals with parameters as BMI and TG/apoB emphasises the potential impact of a perturbed TG metabolism and lifestyle in development of their dyslipidaemia. Concerning the mixed cluster, there was a negative association with PC2, PC3 and PC4, as well as the occurrence of lower values comparatively to other clusters of the parameters that most contribute to this cluster. These findings might be explained by the presence of some FH+ individuals that decrease the levels of parameters associated to TG and reverse cholesterol pathways, while the simultaneous presence of a considerable number of FH- individuals decreases the levels of parameters involved in LDL/apoB pathway.

The application of the Imp_B model in the "All" subset allowed us to acquire the probability of each individual belonging to FH+ and FH- class. The consequent identification of borderline individuals, with higher prevalence among the mixed cluster, support the hypothesis that this

cluster represents a mixed phenotype of dyslipidaemia. Accordingly, the mixed cluster presents a milder lipid profile in comparison to the individuals of the predominant FH+ individuals, but a more severe lipid profile in relation to individuals of the predominant FH- cluster. Further, evidence may suggest that the FH+ individuals of the mixed cluster have a milder profile than the FH+ individuals at their left side of the dendrogram, while the FH- individuals of the mixed individuals present a more severe profile rather than FH- subjects present at their right.

In summary, the hierarchical clustering analysis allowed the identification of biochemical, molecular, and anthropomorphic patterns among individuals, beyond their classification as FH+ and FH-, which contributed to the identification of potential biomarkers and pathways of interest within lipid metabolism. In this way, these findings might help at improving the distinction between individuals and contribute to the understanding of dyslipidaemic mechanisms beyond the lipid profile of FH- individuals.

3. Creation of a new lipid knowledge base directed to dyslipidaemia

Considering the literature review and the potential biomarkers previously identified during modelling and clustering analysis, an oriented search for metabolic pathways of interest allowed to establish a target list of genes. Accordingly, this search resulted in the selection of 14 metabolic pathways whose genes were compiled in a single list – known as target genes. Some of the previously identified biomarkers are well known players in lipid metabolism, and most specifically in FH, including TC and LDL-C [10]. Other biomarkers that revealed to have an important contribution to distinguish FH+ and FH- individuals were TG, TG/apoB, apoB/apoA1, TC/HDL-C, HDL-C, IDL, VLDL-C and LDL-C subfractions (mainly LDL1). This recognizes the important role of VLDL-IDL-LDL delipidation cascade and the different pathways involved in lipoprotein metabolism to understand the biological context under dyslipidaemic states. Also, TG metabolism was revealed to be vital for a better distinction between dyslipidaemias. Indeed, high TG levels are mostly linked to FH- individuals and fat-rich diets, being a frequent trace of polygenic dyslipidaemia profiles [35], [67], [149]. The relevance of these biomarkers was highlighted by their association to some of the biological patterns identified on target genes, considering the additional layers of information collected for the target list and that was integrated in a new lipid knowledge base – MylipidgenesKB. This additional information included associated metabolic pathways, gene expression profiles for tissues of interest, associated GWAS traits and GO terms (for BP and MF domains). Considering the nine selected GWAS traits associated to target genes, some of them correspond

to the previously identified biomarkers including TC, LDL-C, VLDL-C, HDL-C, TG (through "hypertriglyceridaemia", which means high TG levels), and IDL. Regarding the lipid-specific GO terms associated with target genes, some of them are related to biomarkers like lipoprotein species (VLDL, LDL, HDL), TC (through "cholesterol" single designation), and TG.

In addition, a detailed analysis of the associated information collected for target genes allowed to identify a special group among these genes, known as candidate genes, which presented association with at least one of the selected GWAS traits and to lipid-specific GO terms of both BP and MF domains. These candidate genes might shed further light in the distinction between FH+ and FH- individuals, given the fact that they are potential new targets for molecular studies, including future GWAS. The majority of candidate genes are well-known players in lipid metabolism and/or dyslipidaemia, including the coding genes for apolipoproteins, lipases, and proteins associated with lipid receptors and transport [9].

The new lipid knowledge base provides easy access to a specific list of genes known to be involved in several lipid metabolic pathways and/or with a potentially important role in dyslipidaemia metabolic context, as well as it promotes an integrative analysis of distinct layers of information by presenting gene expression data and metabolic, phenotypic, and functional information in a single place. Thus, MylipidgenesKB contributes to the improvement of the lipid metabolism knowledge base and represents a useful resource for the scientific community.

4. Final remarks

The application of machine-learning methods, in opposition to a traditional approach based on pairwise statistical tests, contributes for an improved selection of biomarkers and thus for a better distinction among dyslipidaemic individuals. This is well represented by the higher performance of SB models in comparison to the established SB cut-offs, using the same biochemical parameters (TC and LDL-C) in the same universe of individuals. Another important contribution for the identification of potential FH biomarkers was the use of an extended lipid profile, combining biochemical parameters commonly used for the clinical diagnosis of FH with other parameters not routinely prescribed by physicians and mostly reserved to research centres. Together, the alternative approach of data analysis and the extended lipid profile made possible two of the three major contributions of this project. The first corresponds to the set of 10 models able to classify FH individuals with higher specificity than currently used clinical criteria while conserving good sensitivity values. The second major contribution was achieved by the hierarchical clustering analysis that revealed a third group of

patients, composed of both FH+ and FH- individuals, yet representing a complex biological background that did not allow a clear characterization of these individuals. Afterwards, the results achieved by modelling and clustering analysis allowed the identification of a group of potential biomarkers, which contributed to establish a list of target genes, which represented the initial step to build the MylipidgenesKB – a lipid knowledge base comprising curated information collected from several public databases and that represent the third major contribution of this project.

In conclusion, the work developed in this project can potentially contribute to a better selection of individuals submitted to FH genetic testing, which is essential to improve the diagnostic yield. Additionally, MylipidgenesKB represents a useful starting resource for anyone interested in exploring the lipid metabolic pathways and looking for expression, phenotypic and functional patterns among target genes, in the context of lipid metabolism and dyslipidaemia.

References

[1] J. L. Jain, *Fundamentals of Biochemistry*, Sixth edition, S. Chand & Company LTD., 2005.

[2] D. L. Nelson and M. M. Cox, *Lehninger Principles of Biochemistry*, Fourth edition, W. H. Freeman, 2004.

[3] A. J. Vander, J. Sherman, and D. S. Luciano, *Human Physiology: The Mechanisms of Body Function*, Eighth edition, McGraw-Hill Education, 2001.

[4] A. C. Guyton and J. E. Hall, *Textbook of Medical Physiology*, Eleventh edition, Saunders Elsevier, 2005.

[5] P. O. Kwiterovich, "Diagnosis and Management of Familial Dyslipoproteinemias," *Curr Cardiol Rep*, vol. 15, no. 6, pp. 371-391, Jun. 2013.

[6] M. W. Freeman, "Lipid Metabolism and Coronary Artery Disease," in *Principles of Molecular Medicine*, Second edition, M. S. Runge and C. Patterson (Eds.), Humana Press Inc., 2006, pp. 130–137.

[7] H. Lodish, A. Berk, P. Matsudaira, *et al.*, *Molecular Cell Biology*, Fifth edition, W. H. Freeman, 2003.

[8] A. Rodriguez-Oquendo and P. O. Kwiterovich, "Dyslipidaemias," in *Inborn Metabolic Diseases*, Fifth edition, J. M. Saudubray, et al. (Eds.), Springer, 2012, pp. 439–460.

[9] R. A. Hegele, "Plasma lipoproteins: genetic influences and clinical implications," *Nat Rev Genet*, vol. 10, no. 2, pp. 109–121, Feb. 2009.

[10] J. C. Defesche, S. S. Gidding, M. Harada-Shiba, et al., "Familial hypercholesterolaemia," *Nat Rev Dis Prim*, vol. 3, no. 17093, pp. 1-20, Dec. 2017.

[11] M. M. Jawi, J. Frohlich, and S. Y. Chan, "Lipoprotein(a) the Insurgent: A New Insight into the Structure, Function, Metabolism, Pathogenicity, and Medications Affecting Lipoprotein(a) Molecule," *J Lipids*, vol. 2020, pp. 1–26, Feb. 2020.

[12] L. Renee Ruhaak, A. van der Laarse, and C. M. Cobbaert, "Apolipoprotein profiling as a personalized approach to the diagnosis and treatment of dyslipidaemia," *Ann Clin Biochem Int J Lab Med*, vol. 56, no. 3, pp. 338–356, May 2019.

[13] I. Ramasamy, "Recent advances in physiological lipoprotein metabolism," *Clin Chem Lab Med*, vol. 52, no. 12, pp. 1695–1727, Jan. 2014.

[14] H. R. Superko, "Advanced Lipoprotein Testing and Subfractionation Are Clinically Useful," *Circulation*, vol. 119, no. 17, pp. 2383–2395, May 2009.

[15] I. Ramasamy, "Update on the molecular biology of dyslipidemias," *Clin Chim Acta*, vol. 454, pp. 143–185, Feb. 2016.

[16] A. Vandermeersch, S. Ameye, D. Puype, et al., "Estimation of the low-density lipoprotein (LDL) subclass phenotype using a direct, automated assay of small dense LDL-cholesterol without sample pretreatment," *Clin Chim Acta*, vol. 411, no. 17–18, pp. 1361–1366, Sep. 2010.

[17] R. C. Maranhão, P. O. Carvalho, C. C. Strunz, et al., "Lipoprotein (a): Structure, Pathophysiology and Clinical Implications," *Arq Bras Cardiol*, vol. 103, no. 1, pp. 76–84, Jul.

2014.

[18] C. Paththinige, N. Sirisena, and V. Dissanayake, “Genetic determinants of inherited susceptibility to hypercholesterolemia – a comprehensive literature review,” *Lipids Health Dis*, vol. 16, no. 103, pp. 1-22, Jun. 2017.

[19] R. W. Mahley and Z. S. Ji, “Remnant lipoprotein metabolism: key pathways involving cell-surface heparan sulfate proteoglycans and apolipoprotein E,” *J Lipid Res*, vol. 40, no. 1, pp. 1–16, Jan. 1999.

[20] R. J. Havel and J. P. Kane, “Structure and Metabolism of Lipoproteins,” in *The metabolic and molecular bases of inherited disease*, Seventh edition, C. Scriver, A. Beaudet, W. Sly, et al. (Eds.), McGraw-Hill, 1995, pp. 1841–1851.

[21] D. B. van Schalkwijk, A. A. de Graaf, B. van Ommen, *et al.*, “Improved cholesterol phenotype analysis by a model relating lipoprotein life cycle processes to particle size,” *J Lipid Res*, vol. 50, no. 12, pp. 2398–2411, Dec. 2009.

[22] X. Chen, L. Zhou, and M. M. Hussain, “Lipids and Dyslipoproteinemia,” in *Henry’s Clinical Diagnosis and Management by Laboratory Methods*, Twenty third edition, R. McPherson and M. Pincus (Eds.), Elsevier, 2016, pp. 221–243.

[23] P. A. S. Alphonse and P. J. H. Jones, “Revisiting Human Cholesterol Synthesis and Absorption: The Reciprocity Paradigm and its Key Regulators,” *Lipids*, vol. 51, no. 5, pp. 519–536, May 2016.

[24] M. S. Afonso, R. M. Machado, M. Lavrador, et al., “Molecular Pathways Underlying Cholesterol Homeostasis,” *Nutrients*, vol. 10, no. 6, pp. 760-778, Jun. 2018.

[25] M. S. Brown, A. Radhakrishnan, and J. L. Goldstein, “Retrospective on Cholesterol Homeostasis: The Central Role of Scap,” *Annu Rev Biochem*, vol. 87, no. 1, pp. 783–807, Jun. 2018.

[26] J. A. Dubland and G. A. Francis, “Lysosomal acid lipase: at the crossroads of normal and atherogenic cholesterol metabolism,” *Front Cell Dev Biol*, vol. 3, no. 3, pp. 1–11, Feb. 2015.

[27] P. W. F. Wilson, R. B. D’Agostino, D. Levy, et al., “Prediction of Coronary Heart Disease Using Risk Factor Categories,” *Circulation*, vol. 97, no. 18, pp. 1837–1847, May 1998.

[28] S. Hirayama and T. Miida, “Small dense LDL: An emerging risk factor for cardiovascular disease,” *Clin Chim Acta*, vol. 414, pp. 215–224, Dec. 2012.

[29] J. Vekic, A. Zeljkovic, Z. Jelic-Ivanovic, *et al.*, “Small, dense LDL cholesterol and apolipoprotein B: Relationship with serum lipids and LDL size,” *Atherosclerosis*, vol. 207, no. 2, pp. 496–501, Dec. 2009.

[30] N. Clouet-Foraison, F. Gaie-Levrel, P. Gillery, et al., “Advanced lipoprotein testing for cardiovascular diseases risk assessment: a review of the novel approaches in lipoprotein profiling,” *Clin Chem Lab Med*, vol. 55, no. 10, pp. 1453–1464, May 2017.

[31] F. M. Sacks, “The crucial roles of apolipoproteins E and C-III in apoB lipoprotein metabolism

in normolipidemia and hypertriglyceridemia," *Curr Opin Lipidol*, vol. 26, no. 1, pp. 56–63, Feb. 2015.

[32] Y. Fukushima, S. Hirayama, T. Ueno, *et al.*, "Small dense LDL cholesterol is a robust therapeutic marker of statin treatment in patients with acute coronary syndrome and metabolic syndrome," *Clin Chim Acta*, vol. 412, no. 15–16, pp. 1423–1427, Jul. 2011.

[33] A. J. Berberich and R. A. Hegele, "The complex molecular genetics of familial hypercholesterolaemia," *Nat Rev Cardiol*, vol. 16, no. 1, pp. 9–20, Jan. 2019.

[34] K. J. Moore and M. W. Freeman, "Scavenger receptors in atherosclerosis: Beyond lipid uptake," *Arterioscler Thromb Vasc Biol*, vol. 26, no. 8, pp. 1702–1711, Aug. 2006.

[35] A. M. Medeiros, A. C. Alves, P. Aguiar, et al., "Cardiovascular risk assessment of dyslipidemic children: analysis of biomarkers to identify monogenic dyslipidemia," *J Lipid Res*, vol. 55, no. 5, pp. 947–955, May 2014.

[36] P. J. Talmud, S. Shah, R. Whittall, *et al.*, "Use of low-density lipoprotein cholesterol gene score to distinguish patients with polygenic and monogenic familial hypercholesterolaemia: a case-control study," *Lancet*, vol. 381, no. 9874, pp. 1293–1301, Apr. 2013.

[37] H. M. Kingston, *ABC of Clinical Genetics (ABC Series)*, Third edition, BMJ Books, 2002.

[38] M. Bourbon, Q. Rato, and Investigadores do Estudo Português de Hipercolesterolemia Familiar, "Portuguese Familial Hypercholesterolemia Study: presentation of the study and preliminary results.," *Rev Port Cardiol*, vol. 25, no. 11, pp. 999–1013, Nov. 2006.

[39] D. S. FREDRICKSON and R. S. LEES, "A System for Phenotyping Hyperlipoproteinemia," *Circulation*, vol. 31, no. 3, pp. 321–327, Mar. 1965.

[40] J. L. Beaumont, L. A. Carlson, G. R. Cooper, et al., "Classification of hyperlipidaemias and hyperlipoproteinaemias.," *Bull World Health Organ*, vol. 43, no. 6, pp. 891–915, 1970.

[41] A. J. Brahm and R. A. Hegele, "Chylomicronaemia—current diagnosis and future therapies," *Nat Rev Endocrinol*, vol. 11, no. 6, pp. 352–362, Jun. 2015.

[42] A. Chait and S. Subramanian, "Hypertriglyceridemia: Pathophysiology, Role of Genetics, Consequences, and Treatment," in *Endotext*, K. R. Feingold, B. Anawalt, A. Boyce, et al. (Eds.), MDText.com, Inc., 2000.

[43] M. D. Shapiro, "Rare Genetic Disorders Altering Lipoproteins," in *Endotext*, K. R. Feingold, B. Anawalt, A. Boyce, et al. (Eds.), MDText.com, Inc., 2000.

[44] L. E. Akioyamen, J. Genest, S. Shan, *et al.*, "Estimating the prevalence of heterozygous familial hypercholesterolaemia: a systematic review and meta-analysis," *BMJ Open*, vol. 7, no. 9, p. e016461, Sep. 2017.

[45] M. P. McGowan, S. H. Hosseini Dehkordi, P. M. Moriarty, et al., "Diagnosis and Treatment of Heterozygous Familial Hypercholesterolemia," *J Am Heart Assoc*, vol. 8, no. 24, p. e013225, Dec. 2019.

[46] A. J. Vallejo-Vaz, M. de Marco, C. Stevens, *et al.*, "Overview of the current status of familial

hypercholesterolaemia care in over 60 countries - The EAS Familial Hypercholesterolaemia Studies Collaboration (FHSC)," *Atherosclerosis*, vol. 277, pp. 234–255, Oct. 2018.

[47] M. Bourbon, A. C. Alves, A. M. Medeiros, et al., "Familial hypercholesterolaemia in Portugal," *Atherosclerosis*, vol. 196, no. 2, pp. 633–642, Feb. 2008.

[48] A. M. Medeiros, A. C. Alves, V. Francisco, et al., "Update of the Portuguese Familial Hypercholesterolaemia Study," *Atherosclerosis*, vol. 212, no. 2, pp. 553–558, Oct. 2010.

[49] A. Benito-Vicente, A. C. Alves, A. Etxebarria, et al., "The importance of an integrated analysis of clinical, molecular, and functional data for the genetic diagnosis of familial hypercholesterolemia," *Genet.Med.*, vol. 17, no. 12, pp. 980-988, Mar. 2015.

[50] A. C. Alves, A. Benito-Vicente, A. M. Medeiros, et al., "Further evidence of novel APOB mutations as a cause of familial hypercholesterolaemia," *Atherosclerosis*, vol. 277, pp. 448–456, Oct. 2018.

[51] A. M. Medeiros, A. C. Alves, and M. Bourbon, "Mutational analysis of a cohort with clinical diagnosis of familial hypercholesterolemia: considerations for genetic diagnosis improvement," *Genet Med*, vol. 18, no. 4, pp. 316–324, Apr. 2016.

[52] J. R. Chora, A. M. Medeiros, A. C. Alves, et al., "Analysis of publicly available LDLR, APOB, and PCSK9 variants associated with familial hypercholesterolemia: application of ACMG guidelines and implications for familial hypercholesterolemia diagnosis," *Genet Med*, vol. 20, no. 6, pp. 591–598, Jun. 2018.

[53] A. Sarraju and J. W. Knowles, "Genetic Testing and Risk Scores: Impact on Familial Hypercholesterolemia," *Front Cardiovasc Med*, vol. 6, no. 5, pp. 1-7, Jan. 2019.

[54] A. Benito-Vicente, K. Uribe, S. Jebari, et al., "Familial Hypercholesterolemia: The Most Frequent Cholesterol Metabolism Disorder Caused Disease," *Int J Mol Sci*, vol. 19, no. 11, pp. 3426-3447, Nov. 2018.

[55] A. C. Alves, A. Etxebarria, A. M. Medeiros, *et al.*, "Characterization of the First PCSK9 Gain of Function Homozygote," *J Am Coll Cardiol*, vol. 66, no. 19, pp. 2152–2154, Nov. 2015.

[56] A. Benito-Vicente, K. Uribe, S. Jebari, et al., "Validation of LDLr Activity as a Tool to Improve Genetic Diagnosis of Familial Hypercholesterolemia: A Retrospective on Functional Characterization of LDLr Variants," *Int J Mol Sci*, vol. 19, no. 6, pp. 1676-1694, Jun. 2018.

[57] M. A. Iacocca, J. R. Chora, A. Carrié, *et al.*, "ClinVar database of global familial hypercholesterolemia-associated DNA variants," *Hum Mutat*, vol. 39, no. 11, pp. 1631–1640, Nov. 2018.

[58] A. J. Hooper, J. R. Burnett, D. A. Bell, et al., "The Present and the Future of Genetic Testing in Familial Hypercholesterolemia: Opportunities and Caveats," *Curr Atheroscler Rep*, vol. 20, no. 31, pp. 1-7, Jun. 2018.

[59] M. do C. Espinheira, C. Vasconcelos, A. M. Medeiros, et al., "Hypercholesterolemia – A disease with expression since childhood," *Rev Port Cardiol (English Ed*, vol. 32, no. 5, pp.

379–386, May 2013.

[60] Scientific Steering Committee on behalf of the Simon Broome Register Group, “Risk of fatal coronary heart disease in familial hypercholesterolaemia,” *BMJ*, vol. 303, no. 6807, pp. 893–896, Oct. 1991.

[61] A. C. Sturm, J. W. Knowles, S. S. Gidding, *et al.*, “Clinical Genetic Testing for Familial Hypercholesterolemia,” *J Am Coll Cardiol*, vol. 72, no. 6, pp. 662–680, Aug. 2018.

[62] F. Mach, C. Baigent, A. Catapano, *et al.*, “2019 ESC/EAS Guidelines for the management of dyslipidaemias: lipid modification to reduce cardiovascular risk,” *Eur Heart J*, vol. 00, pp. 1–78, Aug. 2019.

[63] A. Wiegman, S. S. Gidding, G. F. Watts, *et al.*, “Familial hypercholesterolaemia in children and adolescents: gaining decades of life by optimizing detection and treatment,” *Eur Heart J*, vol. 36, no. 36, pp. 2425–2437, Sep. 2015.

[64] B. Okopień, Ł. Bułdak, and A. Bołdys, “Current and future trends in the lipid lowering therapy,” *Pharmacol Reports*, vol. 68, no. 4, pp. 737–747, Aug. 2016.

[65] J. L. Ross, “Statins in the Management of Pediatric Dyslipidemia,” *J Pediatr Nurs*, vol. 31, no. 6, pp. 723–735, Nov. 2016.

[66] Y. Ghaleb, S. Elbitar, P. El Khoury, *et al.*, “Usefulness of the genetic risk score to identify phenocopies in families with familial hypercholesterolemia?,” *Eur J Hum Genet*, vol. 26, no. 4, pp. 570–578, Apr. 2018.

[67] M. Futema, M. Bourbon, M. Williams, et al., “Clinical utility of the polygenic LDL-C SNP score in familial hypercholesterolemia,” *Atherosclerosis*, vol. 277, pp. 457–463, Oct. 2018.

[68] B. Sjouke, M. Tanck, S. Fouchier, *et al.*, “Children with hypercholesterolemia of unknown cause: Value of genetic risk scores,” *J Clin Lipidol*, vol. 10, no. 4, pp. 851–859, Jul. 2016.

[69] C. Mariano, A. C. Alves, A. M. Medeiros, *et al.*, “The familial hypercholesterolaemia phenotype: Monogenic familial hypercholesterolaemia, polygenic hypercholesterolaemia and other causes,” *Clin Genet*, vol. 97, no. 3, pp. 457–466, Mar. 2020.

[70] A. Vuorio, G. F. Watts, and P. T. Kovanen, “Lipoprotein(a) as a risk factor for calcific aortic valvulopathy in heterozygous familial hypercholesterolemia,” *Atherosclerosis*, vol. 281, pp. 25–30, Feb. 2019.

[71] C. Andersson and R. S. Vasan, “Epidemiology of cardiovascular disease in young individuals,” *Nat Rev Cardiol*, vol. 15, no. 4, pp. 230–240, Apr. 2018.

[72] B. A. Goldstein, A. M. Navar, and R. E. Carter, “Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges,” *Eur Heart J*, vol. 38, no. 23, pp. 1805–1814, Jul. 2016.

[73] S. F. Weng, J. Reps, J. Kai, et al., “Can machine-learning improve cardiovascular risk prediction using routine clinical data?,” *PLoS One*, vol. 12, no. 4, p. e0174944, Apr. 2017.

[74] R. K. Sevakula, W. M. Au-Yeung, J. P. Singh, et al., “State-of-the-Art Machine Learning

Techniques Aiming to Improve Patient Outcomes Pertaining to the Cardiovascular System," *J Am Heart Assoc*, vol. 9, no. 4, p. e013924, Feb. 2020.

[75] A. Lamaziere, C. Wolf, and P. J. Quinn, "Perturbations of Lipid Metabolism Indexed by Lipidomic Biomarkers," *Metabolites*, vol. 2, no. 4, pp. 1–18, Jan. 2012.

[76] M. A. Iacocca and R. A. Hegele, "Recent advances in genetic testing for familial hypercholesterolemia," *Expert Rev Mol Diagn*, vol. 17, no. 7, pp. 641–651, Jul. 2017.

[77] J. A. Reuter, D. V. Spacek, and M. P. Snyder, "High-Throughput Sequencing Technologies," *Mol Cell*, vol. 58, no. 4, pp. 586–597, May 2015.

[78] C. T. Johansen, J. B. Dubé, M. N. Loyzer, *et al.*, "LipidSeq: a next-generation clinical resequencing panel for monogenic dyslipidemias," *J Lipid Res*, vol. 55, no. 4, pp. 765–772, Apr. 2014.

[79] A. V. Khera, M. Chaffin, S. M. Zekavat, *et al.*, "Whole-Genome Sequencing to Characterize Monogenic and Polygenic Contributions in Patients Hospitalized With Early-Onset Myocardial Infarction," *Circulation*, vol. 139, no. 13, pp. 1593–1602, Mar. 2019.

[80] M. Trinder, M. Paquette, L. Cermakova, *et al.*, "Polygenic Contribution to Low-Density Lipoprotein Cholesterol Levels and Cardiovascular Risk in Monogenic Familial Hypercholesterolemia," *Circ Genomic Precis Med*, vol. 13, no. 5, pp. 515–523, Oct. 2020.

[81] A. V. Khera, M. Chaffin, K. G. Aragam, *et al.*, "Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations," *Nat Genet*, vol. 50, no. 9, pp. 1219–1224, Sep. 2018.

[82] T. J. Hoffmann, E. Theusch, T. Haldar, *et al.*, "A large electronic-health-record-based genome-wide study of serum lipids," *Nat Genet*, vol. 50, no. 3, pp. 401–413, Mar. 2018.

[83] S. W. van der Laan, E. L. Harshfield, D. Hemerich, et al., "From lipid locus to drug target through human genomics," *Cardiovasc Res*, vol. 114, pp. 1258–1270, May 2018.

[84] M. Buscot, C. G. Magnussen, M. Juonala, *et al.*, "The Combined Effect of Common Genetic Risk Variants on Circulating Lipoproteins Is Evident in Childhood: A Longitudinal Analysis of the Cardiovascular Risk in Young Finns Study," *PLoS One*, vol. 11, no. 1, p. e0146081, Jan. 2016.

[85] R. Tabassum, J. T. Rämö, P. Ripatti, *et al.*, "Genetic architecture of human plasma lipidome and its link to cardiovascular disease," *Nat Commun*, vol. 10, no. 1, p. 4329, Dec. 2019.

[86] A. J. Cupido, T. R. Tromp, and G. K. Hovingh, "The clinical applicability of polygenic risk scores for LDL-cholesterol: considerations, current evidence and future perspectives," *Curr Opin Lipidol*, vol. 32, no. 2, pp. 112–116, Apr. 2021.

[87] W. J. Murdoch, C. Singh, K. Kumbier, et al., "Definitions, methods, and applications in interpretable machine learning," *Proc Natl Acad Sci*, vol. 116, no. 44, pp. 22071–22080, Oct. 2019.

[88] G. James, D. Witten, T. Hastie, et al., *An Introduction to Statistical Learning*, First edition,

Springer, 2013.

[89] M. Kuhn and K. Johnson, *Applied Predictive Modeling*, First edition, Springer, 2013.

[90] A. Kassambara, *Practical Guide to Cluster Analysis in R: Unsupervised machine learning*, First edition, STHDA, 2017.

[91] D. R. Edla, D. Tripathi, V. Kuppili, et al., "Survey on Clustering Techniques," in *Proceedings of the 2nd International Conference on Inventive Communication and Computational Technologies*, 2018, pp. 696–703.

[92] P. D'urso and L. De Giovanni, "Unsupervised Learning," in *Wiley Encyclopedia of Electrical and Electronics Engineering*, J. Webster (Ed.), John Wiley & Sons, Inc, 2018, pp. 1–23.

[93] M. P. H. Stumpf, D. J. Balding, and M. Girolami (Eds.), *Handbook of Statistical Systems Biology*, First Edition, John Wiley & Sons Ltd, 2011.

[94] A. O'Connor, C. J. Brasher, D. A. Slatter, *et al.*, "LipidFinder: A computational workflow for discovery of lipids identifies eicosanoid-phosphoinositides in platelets," *JCI Insight*, vol. 2, no. 7, Apr. 2017.

[95] E. Fahy, M. Sud, D. Cotter, et al., "LIPID MAPS online tools for lipid research," *Nucleic Acids Res*, vol. 35, no. suppl_2, pp. W606–W612, May 2007.

[96] J. M. Foster, P. Moreno, A. Fabregat, *et al.*, "LipidHome: A Database of Theoretical Lipids Optimized for High Throughput Mass Spectrometry Lipidomics," *PLoS One*, vol. 8, no. 5, p. e61951, May 2013.

[97] B. Kormeier, K. Hippe, T. Töpel, et al., "CardioVINEdb: a data warehouse approach for integration of life science data in cardiovascular diseases.," *J Integr Bioinform*, vol. 7, no. 1, p. 142, 2010.

[98] D. M. Hoefner, S. D. Hodel, J. F. O'Brien, *et al.*, "Development of a rapid, quantitative method for LDL subfractionation with use of the Quantimetrix Lipoprint LDL System.," *Clin Chem*, vol. 47, no. 2, pp. 266–74, Feb. 2001.

[99] WHO MULTICENTRE GROWTH REFERENCE STUDY GROUP, "WHO Child Growth Standards based on length/height, weight and age," *Acta Pædiatrica*, vol. 95, pp. 76–85, 2006.

[100] M. de Onis, "Development of a WHO growth reference for school-aged children and adolescents," *Bull World Health Organ*, vol. 85, no. 9, pp. 660–667, Sep. 2007.

[101] R Core Team, "R: A language and environment for statistical computing," *R Foundation for Statistical Computing*. Vienna, 2017.

[102] M. K. C. from Jed Wing *et al.*, "caret: Classification and Regression Training." 2018.

[103] A. Gelman and Y. S. Su, "arm: Data Analysis Using Regression and Multilevel/Hierarchical Models." 2018.

[104] A. J. Bishara and J. B. Hittner, "Testing the significance of a correlation with nonnormal data: Comparison of Pearson, Spearman, transformation, and resampling approaches.," *Psychol Methods*, vol. 17, no. 3, pp. 399–417, 2012.

[105] D. Robinson and A. Hayes, “broom: Convert Statistical Analysis Objects into Tidy Tibbles.” 2018.

[106] X. Robin *et al.*, “pROC: an open-source package for R and S+ to analyze and compare ROC curves,” *BMC Bioinformatics*, vol. 12, p. 77, 2011.

[107] J. R. Landis and G. G. Koch, “The Measurement of Observer Agreement for Categorical Data,” *Biometrics*, vol. 33, no. 1, pp. 159–174, Mar. 1977.

[108] S. Lê, J. Josse, and F. Husson, “FactoMineR : An R Package for Multivariate Analysis,” *J Stat Softw*, vol. 25, no. 1, pp. 1-18, Mar. 2008.

[109] F. Husson, “Principal Component Analysis (PCA),” *R documentation*, 2018. [Online]. Available: https://www.rdocumentation.org/packages/FactoMineR/versions/1.41/topics/PCA. [Accessed: 13-Aug-2018].

[110] F. Husson, A. Julie, and J. Pagès, “Principal component methods -hierarchical clustering - partitional clustering: why would we need to choose for visualizing data?,” Technical Report of the Applied Mathematics Department (Agrocampus - Ouest), pp. 1-17, Sep. 2010.

[111] F. Husson, “Hierarchical Clustering on Principle Components (HCPC),” *R documentation*, 2018. [Online]. Available: https://www.rdocumentation.org/packages/FactoMineR/versions/1.41/topics/HCPC. [Accessed: 13-Aug-2018].

[112] A. Kassambara and F. Mundt, “factoextra: Extract and Visualize the Results of Multivariate Data Analyses.” 2017.

[113] A. Kassambara, “factoextra: Extract and Visualize the Results of Multivariate Data Analyses,” *R documentation*, 2017. [Online]. Available: https://www.rdocumentation.org/packages/factoextra. [Accessed: 13-Aug-2018].

[114] H. Wickham, R. François, L. Henry, and K. Müller, “dplyr: A Grammar of Data Manipulation.” 2019.

[115] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*, Second edition, Springer, 2016.

[116] A. R. Pico, T. Kelder, M. P. van Iersel, et al., “WikiPathways: Pathway Editing for the People,” *PLoS Biol*, vol. 6, no. 7, p. e184, Jul. 2008.

[117] D. W. Huang, B. T. Sherman, and R. A. Lempicki, “Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources,” *Nat Protoc*, vol. 4, no. 1, pp. 44–57, Jan. 2009.

[118] D. W. Huang, B. T. Sherman, and R. A. Lempicki, “Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists,” *Nucleic Acids Res*, vol. 37, no. 1, pp. 1–13, Jan. 2009.

[119] J. Lonsdale, J. Thomas, M. Salvatore, *et al.*, “The Genotype-Tissue Expression (GTEx) project,” *Nat Genet*, vol. 45, no. 6, pp. 580–585, Jun. 2013.

[120] A. Buniello, J. A. L. MacArthur, M. Cerezo, *et al.*, “The NHGRI-EBI GWAS Catalog of

published genome-wide association studies, targeted arrays and summary statistics 2019," *Nucleic Acids Res*, vol. 47, no. D1, pp. D1005–D1012, Jan. 2019.

[121] R. Magno and A. T. Maia, "gwasrapidd: an R package to query, download and wrangle GWAS catalog data," *Bioinformatics*, vol. 36, no. 2, pp. 649–650, Aug. 2019.

[122] C. Sievert, *Interactive Web-Based Data Visualization with R, plotly, and shiny*, First edition, Chapman and Hall/CRC Press, 2020.

[123] S. Durinck, P. T. Spellman, E. Birney, et al., "Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt," *Nat Protoc*, vol. 4, no. 8, pp. 1184–1191, Aug. 2009.

[124] S. Durinck, Y. Moreau, A. Kasprzyk, *et al.*, "BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis" *Bioinformatics*, vol. 21, no. 16, pp. 3439–3440, Aug. 2005.

[125] R Core Team, "R: A Language and Environment for Statistical Computing." Vienna, Austria, 2020.

[126] W. Chang, J. Cheng, J. J. Allaire, et al., "shiny: Web Application Framework for R." 2020.

[127] W. Chang and B. Borges Ribeiro, "shinydashboard: Create Dashboards with 'Shiny.'" 2018.

[128] D. Granjon, "shinydashboardPlus: Add More 'AdminLTE2' Components to 'shinydashboard.' " 2020.

[129] V. Perrier, F. Meyer, and D. Granjon, "shinyWidgets: Custom Inputs Widgets for Shiny." 2021.

[130] Y. Xie, J. Cheng, and X. Tan, "DT: A Wrapper of the JavaScript Library 'DataTables.'" 2021.

[131] P. Shannon, "Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks," *Genome Res*, vol. 13, no. 11, pp. 2498–2504, Nov. 2003.

[132] J. Montojo, K. Zuberi, H. Rodríguez, *et al.*, "GeneMANIA Cytoscape plugin: fast gene function predictions on the desktop," *Bioinformatics*, vol. 26, no. 22, pp. 2927–2928, Nov. 2010.

[133] B. Li, A. Sharma, J. Meng, et al., "Applying machine learning to identify autistic adults using imitation: An exploratory study," *PLoS One*, vol. 12, no. 8, p. e0182652, Aug. 2017.

[134] R. Salvador, J. Radua, E. J. Canales-Rodríguez, *et al.*, "Evaluation of machine learning algorithms and structural features for optimal MRI-based diagnostic prediction in psychosis," *PLoS One*, vol. 12, no. 4, p. e0175683, Apr. 2017.

[135] L. Gao, M. Ye, and C. Wu, "Cancer Classification Based on Support Vector Machine Optimized by Particle Swarm Optimization and Artificial Bee Colony," *Molecules*, vol. 22, no. 12, p. 2086, Nov. 2017.

[136] L.-G. Li, X. Yin, and T. Zhang, "Tracking antibiotic resistance gene pollution from different sources using machine-learning classification," *Microbiome*, vol. 6, no. 1, p. 93, Dec. 2018.

[137] M. A. Eissa, N. L. Mihalopoulos, R. Holubkov, et al., "Changes in Fasting Lipids during

Puberty," *J Pediatr*, vol. 170, pp. 199–205, Mar. 2016.

[138] T. A. Lagace, "PCSK9 and LDLR degradation," *Curr Opin Lipidol*, vol. 25, no. 5, pp. 387–393, Oct. 2014.

[139] C. Beck, "Assembly and Secretion of Atherogenic Lipoproteins," Göteborg University. Sahlgrenska Academy, 2008.

[140] G. S. Sagoo, I. Tatt, G. Salanti, *et al.*, "Seven Lipoprotein Lipase Gene Polymorphisms, Lipid Fractions, and Coronary Disease: A HuGE Association Review and Meta-Analysis," *Am J Epidemiol*, vol. 168, no. 11, pp. 1233–1246, Oct. 2008.

[141] B. Teng, A. D. Sniderman, A. K. Soutar, et al., "Metabolic basis of hyperapobetalipoproteinemia. Turnover of apolipoprotein B in low density lipoprotein and its precursors and subfractions compared with normal and familial hypercholesterolemia.," *J Clin Invest*, vol. 77, no. 3, pp. 663–672, Mar. 1986.

[142] A. Zambon, S. S. Deeb, J. E. Hokanson, et al., "Common Variants in the Promoter of the Hepatic Lipase Gene Are Associated With Lower Levels of Hepatic Lipase Activity, Buoyant LDL, and Higher HDL 2 Cholesterol," *Arterioscler Thromb Vasc Biol*, vol. 18, no. 11, pp. 1723–1729, Nov. 1998.

[143] T. Hirano, Y. Ito, H. Saegusa, et al., "A novel and simple method for quantification of small, dense LDL," *J Lipid Res*, vol. 44, no. 11, pp. 2193–2201, Nov. 2003.

[144] B. Jassal, L. Matthews, G. Viteri, *et al.*, "The reactome pathway knowledgebase," *Nucleic Acids Res*, vol. 48, no. D1, pp. D498–D503, Jan. 2020.

[145] I. De Castro-Orós, M. Pocoví, and F. Civeira, "The fine line between familial and polygenic hypercholesterolemia," *Clin Lipidol*, vol. 8, no. 3, pp. 303–306, Jun. 2013.

[146] E. Ajufo and M. Cuchel, "Improving the yield of genetic testing in familial hypercholesterolaemia," *Eur Heart J*, vol. 38, pp. 574–576, May 2016.

[147] A. V. Khera, H. Won, G. Peloso, *et al.*, "Diagnostic Yield and Clinical Utility of Sequencing Familial Hypercholesterolemia Genes in Patients With Severe Hypercholesterolemia," *J Am Coll Cardiol*, vol. 67, no. 22, pp. 2578–2589, Jun. 2016.

[148] B. G. Nordestgaard, M. R. Langlois, A. Langsted, *et al.*, "Quantifying atherogenic lipoproteins for lipid-lowering strategies: Consensus-based recommendations from EAS and EFLM," *Atherosclerosis*, vol. 294, pp. 46–61, Feb. 2020.

[149] M. Futema, S. Shah, J. A. Cooper, *et al.*, "Refinement of Variant Selection for the LDL Cholesterol Genetic Risk Score in the Diagnosis of the Polygenic Form of Clinical Familial Hypercholesterolemia and Replication in Samples from 6 Countries," *Clin Chem*, vol. 61, no. 1, pp. 231–238, Jan. 2015.

[150] I. Lamiquiz-Moneo, M. R. Pérez-Ruiz, E. Jarauta, *et al.*, "Single Nucleotide Variants Associated With Polygenic Hypercholesterolemia in Families Diagnosed Clinically With Familial Hypercholesterolemia," *Rev Española Cardiol (English Ed*, vol. 71, no. 5, pp. 351–

356, May 2018.

[151] S. P. Guay, D. Brisson, B. Lamarche, et al., “Epipolymorphisms within lipoprotein genes contribute independently to plasma lipid levels in familial hypercholesterolemia,” *Epigenetics*, vol. 9, no. 5, pp. 718–729, May 2014.

[152] J. D. Brunzell, A. Zambon, and S. S. Deeb, “The effect of hepatic lipase on coronary artery disease in humans is influenced by the underlying lipoprotein phenotype,” *Biochim Biophys Acta - Mol Cell Biol Lipids*, vol. 1821, no. 3, pp. 365–372, Mar. 2012.

[153] T. Hayashi, S. Koba, Y. Ito, et al., “Method for estimating high sdLDL-C by measuring triglyceride and apolipoprotein B levels,” *Lipids Health Dis*, vol. 16, no. 1, p. 21, Dec. 2017.

Annex

Annex 1| *PFHS-ped dataset*

The full PFHS-ped dataset composed by 211 individuals and presenting all the biochemical parameters that correspond to the extended lipid profile (including ratios), gender, age, BMI, class (FH+/FH-), and the affected gene in FH+ individuals. This file can be accessed in the following repository: https://github.com/GamaPintoLab/MartaCorreia-PhD-thesis.git.

Annex 2| *Mean of biochemical parameters by subset and individuals class*

Table A2.1. Mean value of each biochemical parameter by subset, for FH+ individuals. Grey cells correspond to the absence of that parameter in the given subset. sdLDL.Day and sdLDL.Lipo were measured by different techniques, using RX daytona+® analyser and Lipoprint® assay, respectively.

Subsets / Parameters	All	Basic	Advanced	Lipoprint	Basic & Advanced	Basic & Lipoprint	Advanced & Lipoprint
TC	271.32	274.87			270.37	274.47	
LDL-C	203.25	205.66			202.32	205.50	
HDL-C	49.96	51.72			50.42	50.66	
TG	72.86	74.73			71.11	74.09	
Lp(a)	45.21	40.32			43.01	41.94	
ApoB	129.61	129.24			130.63	130.19	
ApoA-I	133.18	135.35			133.47	135.09	
ApoB/ApoA-I	1.01	0.99			1.02	1.00	
TG/ApoB	0.57	0.59			0.56	0.58	
TC/HDL-C	5.74	5.68			5.69	5.75	
ApoA-II	27.54		26.55		26.55		27.54
ApoC-II	3.10		2.94		2.94		3.10
ApoC-III	6.95		6.66		6.66		6.95
ApoE	3.81		3.69		3.69		3.81
sdLDL.Day	41.47		38.44		38.44		41.47
ApoC-II/ApoC-III	0.43		0.43		0.43		0.43
sdLDL/LDL-C	0.20		0.19		0.19		0.20
VLDL	34.39			34.47		34.47	34.39
MIDA	26.46			27.88		27.88	26.46
MIDB	19.71			20.16		20.16	19.71
MIDC	29.18			28.91		28.91	29.18
LDL1	65.07			66.59		66.59	65.07
LDL2	31.96			32.00		32.00	31.96
HDL.Lipo	53.14			53.69		53.69	53.14
sdLDL.Lipo	5.43			5.63		5.63	5.43
IDL	75.36			76.94		76.94	75.36
VLDL/IDL	0.47			0.46		0.46	0.47
VLDL/LDL-C	0.17			0.17		0.17	0.17

Table A2.2. Mean value of each biochemical parameter by subset, for FH- individuals. Grey cells correspond to the absence of that parameter in the given subset. sdLDL.Day and sdLDL.Lipo were measured by different techniques, using RX daytona+® analyser and Lipoprint® assay, respectively.

Subsets / Parameters	All	Basic	Advanced	Lipoprint	Basic & Advanced	Basic & Lipoprint	Advanced & Lipoprint
TC	230.46	228.49			232.36	232.02	
LDL-C	152.96	150.87			155.19	154.86	
HDL-C	60.56	60.11			60.36	59.73	
TG	96.66	93.64			92.78	97.94	
Lp(a)	45.74	54.89			48.93	49.84	
ApoB	96.96	99.33			99.27	98.63	
ApoA-I	164.08	156.69			161.93	162.25	
ApoB/ApoA-I	0.61	0.65			0.64	0.63	
TG/ApoB	1.00	0.95			0.94	0.98	
TC/HDL-C	4.02	3.99			4.06	4.08	
ApoA-II	31.37		30.44		30.44		31.37
ApoC-II	4.52		4.34		4.34		4.52
ApoC-III	9.09		8.82		8.82		9.09
ApoE	3.45		3.37		3.37		3.45
sdLDL.Day	30.97		30.52		30.52		30.97
ApoC-II/ApoC-III	0.51		0.50		0.50		0.51
sdLDL/LDL-C	0.20		0.19		0.19		0.20
VLDL	29.80			30.86		30.86	29.80
MIDA	22.52			23.62		23.62	22.52
MIDB	16.36			17.38		17.38	16.36
MIDC	21.28			21.97		21.97	21.28
LDL1	48.04			47.63		47.63	48.04
LDL2	26.16			24.86		24.86	26.16
HDL.Lipo	57.60			56.94		56.94	57.60
sdLDL.Lipo	7.46			7.67		7.67	7.46
IDL	60.16			62.97		62.97	60.16
VLDL/IDL	0.51			0.51		0.51	0.51
VLDL/LDL-C	0.20			0.20		0.20	0.20

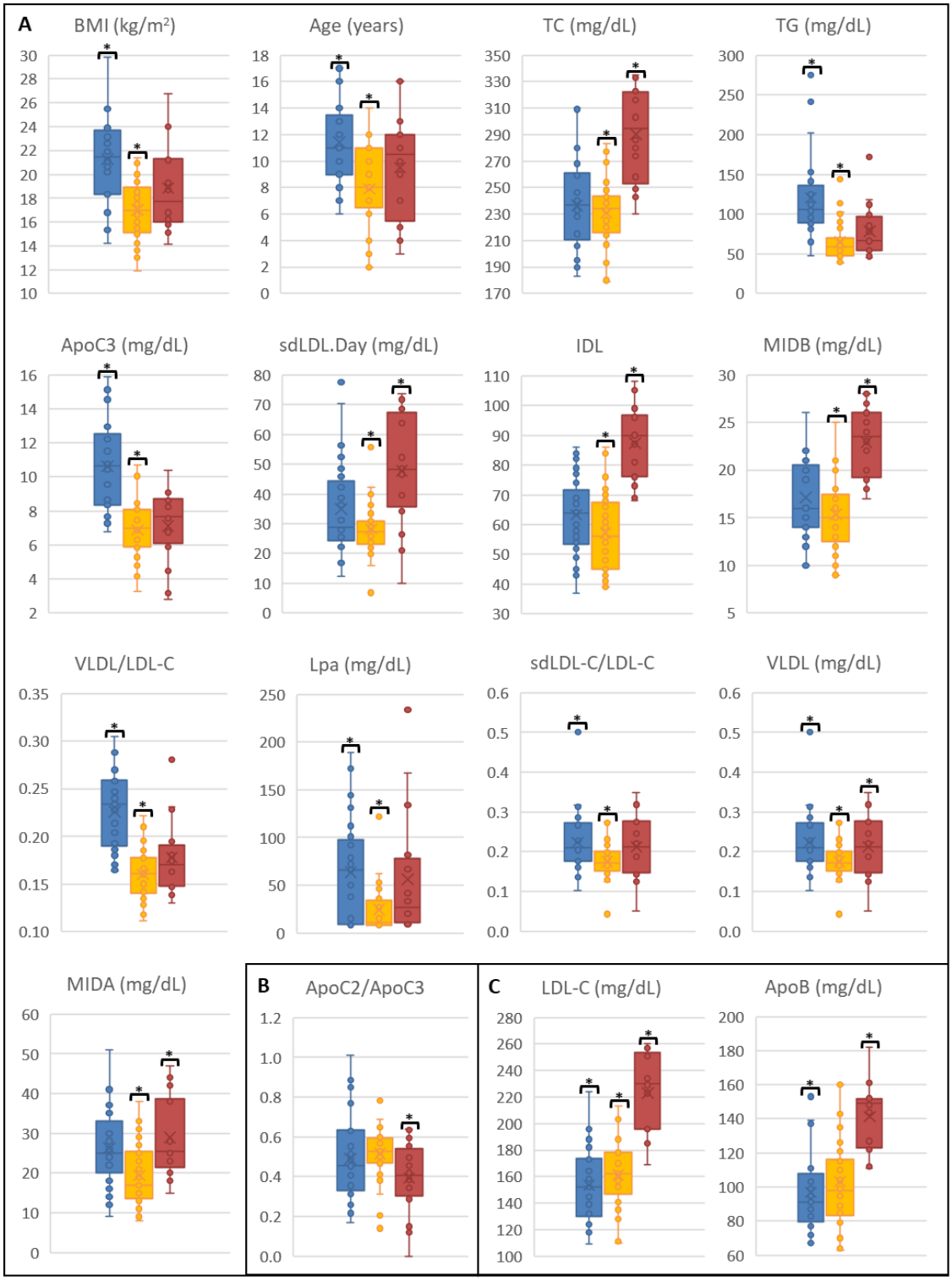
Annex 3| *Full list of the 67 trained classification models*

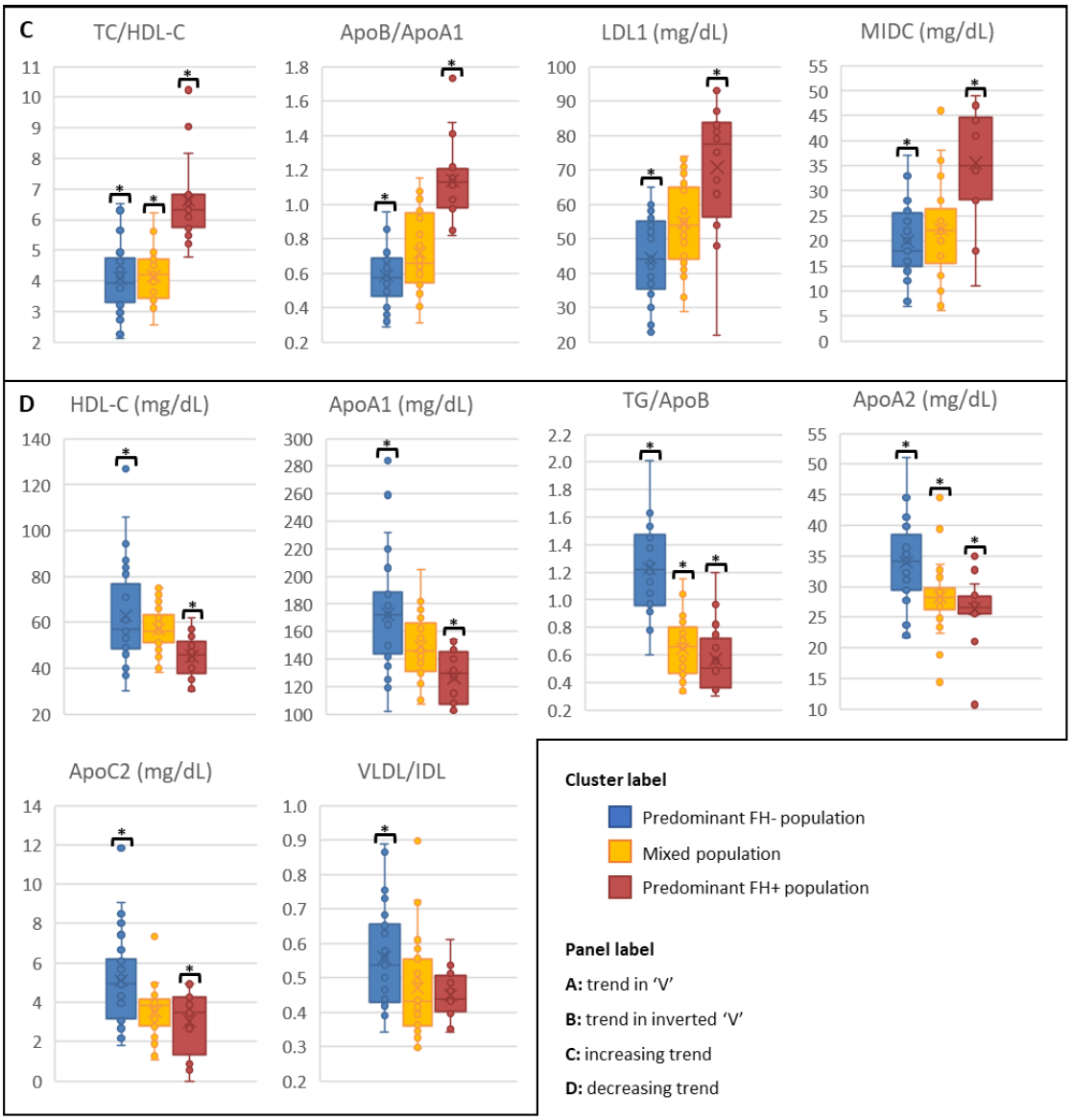
All the trained models, including pilot models, and their performance in the distinction of FH+/FH- individuals. The column titles with green colour fill highlight the statistics used for model ranking. Model names comprise a combination of method and subset designations following the code above: filtering for highly correlated parameters (c of "cor"), selecting only the top parameters of those selected by RFE method (t of "top"), naming the subset used for training the model (A of "All", B of "Basic", Ad of "Advanced", L of "Lipoprint", BAd of "Basic & Advanced", BL of "Basic & Lipoprint", AdL of "Advanced & Lipoprint"). N: number of individuals; Np: number of parameters; Acc: accuracy; k: Cohen's kappa coefficient; Sens: sensitivity; Spec: specificity; TP: number of true positives; FN: number of false negatives; FP: number of false positives; TN: number of true negatives; AUC: area under the ROC curve.

This file can be accessed in the following repository: https://github.com/GamaPintoLab/MartaCorreia-PhD-thesis.git.

Annex 4| *Distribution of the parameters whose difference in the mean was statistically significant at least for one cluster in comparison to the overall mean in the "All" subset*

Figure A4. Distribution of parameters by clusters, with an asterisk showing the clusters whose mean was significantly different from the overall mean, under a confidence level of 95%. Parameters were grouped according to the mean trend among clusters, as present in the panel label. Clusters are identified according to the cluster label.




C
TC/HDL-C
ApoB/ApoA1
LDL1 (mg/dL)
MIDC (mg/dL)
D
HDL-C (mg/dL)
ApoA1 (mg/dL)
TG/ApoB
ApoA2 (mg/dL)
ApoC2 (mg/dL)
VLDL/IDL
Cluster label
Predominant FH- population
Mixed population
Predominant FH+ population
Panel label
A: trend in 'V'
B: trend in inverted 'V'
C: increasing trend
D: decreasing trend

Annex 5| *Categorical variables associated with the PFHS-ped dataset*

The full set of categorical variables available for the "All" subset, which is composed by 78 out of the total 211 individuals from the PFHS-ped dataset. These variables were used as supplementary variables in HCPC analysis and include class (FH+/FH-), gender, SB criteria (yes/no for the fulfilment of TC and LDL-C cut-offs from Simon Broome criteria), BMI class, Lipoprotein profile, activity class (according to the percentage of molecular activity that remains in the affected allele), gene (affected gene), and LDL-C score (according to the LDL-C polygenic risk score). This file can be accessed in the following repository: https://github.com/GamaPintoLab/MartaCorreia-PhD-thesis.git.

Annex 6| *Class probabilities predicted by Imp_B model*

The probability of being classified as FH+ and FH- according to the predictions of Imp_B model, for all the individuals of "All" subset that comprise the work population for cluster characterization. Individuals are listed by the order of appearance in the dendrogram, which means from the predominant FH+ cluster to the predominant FH- cluster. Class and cluster assignments are also present. prob_FH+: probability of being FH+, prob_FH-: probability of being FH-, Δprob: the difference between prob_FH+ and prob_FH-, Δprob_cat: categories of Δprob regarding the ambiguity of classification (clear, reasonable, ambiguous, very ambiguous), Class_pred: predicted classification by Imp_B model, Class_obs: observed classification according to the results of molecular studies.
This file can be accessed in the following repository: https://github.com/GamaPintoLab/MartaCorreia-PhD-thesis.git.

Annex 7| *List of target genes and associated information*

The full list of the 466 target genes with Ensembl ID, gene symbol and full name according to the HGNC database. Gene associated information comprises metabolic pathways, expression profiles for tissues of interest and transcriptome, GWAS traits, and presence or absence of associated lipid-specific GO terms. Exp: (gene) expression, SI: small intestine, Trans: transcriptome, NF: not found (association to any of the selected traits), CAD: coronary artery disease, HTG: hypertriglyceridaemia, CVD: cardiovascular disease, BP: biological process, MF: molecular function.
This file can be accessed in the following repository: https://github.com/GamaPintoLab/MartaCorreia-PhD-thesis.git.

Annex 8| *Full list of lipid-specific GO terms*

Lipid-specific GO terms that were selected as the most representative of target genes, for both GO domains (BP and MF). The terms are grouped according to their hierarchical relations as parent and child terms. Ng: number of associated target genes. This file can be accessed in the following repository: https://github.com/GamaPintoLab/MartaCorreia-PhD-thesis.git.

Annex 9| *Full list of other GO terms*

Other GO terms selected as representative of target genes and that are not related to lipid metabolism. There is a list for each GO domain (BP and MF) and the terms are grouped according to their hierarchical relations as parent and child terms. Ng: number of associated target genes. This file can be accessed in the following repository: https://github.com/GamaPintoLab/MartaCorreia-PhD-thesis.git.