Analysis of Gene Expression and Structural Variations in Common Genes Across Populations: Implications for Cardiovascular Disease and Reverse Cholesterol Pathway

by
Shivansh Verma

Under the supervision of
Dr. Arjun Ray

Submitted in partial fulfilment of the requirements
for the degree of Master of Technology,
Computational Biology

Center for Computational Biology Indraprastha
Institute of Information Technology - Delhi
August 2023

Certificate

This is to certify that the thesis titled *"Analysis of Gene Expression and Structural Variations in Common Genes Across Populations: Implications for Cardiovascular Disease and Reverse Cholesterol Pathway"* being submitted by **Shivansh Verma** to the Indraprastha Institute of Information Technology Delhi, for the award of the Master of Technology is an original research work carried out by him under my supervision. In my opinion, the the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree/diploma.

August,2023

Dr Arjun Ray

Assistant Professor

Department of Computational Biology
Indraprastha Institute of Information Technology Delhi
New Delhi 110 020

Acknowledgements

I am deeply grateful to Prof. Arjun Ray for his unwavering guidance and mentorship throughout this journey. His encouragement and push for excellence have been instrumental in shaping this work. Even during his busiest days, he generously offered his time to help me with any issues and clarify doubts, for which I am truly thankful.

I extend my heartfelt thanks to my mother, brother, and sister, whose unwavering support and understanding allowed me to fully focus on my thesis. Their constant encouragement has been a source of strength throughout my academic journey.

I am indebted to my PhD mentor, Gayatri Panda, Sukriti Sacher and Himanshi Garg, for their invaluable advice and insights that have greatly enriched my thesis work. Her creative suggestions and ideas have played a significant role in shaping the project.

Gratitude also goes to my friends Amit Samal, Anshul Yadav and Dilip Kumar Yadav for their support. Their camaraderie made the ups and downs of the project easier to navigate.I thank my fellow batch mates for their continuous support and camaraderie. Their willingness to help and share knowledge has been a great source of inspiration.

A special mention goes to all the faculty members and staff of the Department of Computational Biology and IIIT Delhi, who have supported us throughout our college journey. I would also like to thank Mr Adarsh from the IT department for his invaluable assistance in providing access to college IT infrastructure and facilitating our web server deployment.

Abstract

Cardiovascular disease (CVD) remains a significant global health challenge, influenced by a complex interplay of genetic and environmental factors. Dysregulated lipid metabolism, a key contributor to CVD, disrupts cholesterol balance and triggers atherosclerotic plaque formation. Understanding how genetics affect lipid balance is crucial for innovative treatments.This study focuses on lipid metabolism, vital for heart health, and its genetic influencers.We combine insights from three independent studies, each uncovering genetic links to lipid metabolism. By merging these findings, we aim for a comprehensive understanding of the genetic landscape.

Initially, we identify important genes from each study and then connect them across diverse populations. These genes hold the key to understanding the genetic basis of CVD.Our research delves into how genetic variations impact protein-ligand interactions in terms of binding energy related to lipid metabolism, using virtual screening. We use a method called molecular docking with Autodock Vina, assessing binding energies among 54,162 compounds. To understand the significance of binding energy differences, we use a statistical test (Kolmogorov-Smirnov or K-S test), revealing significant variations in binding energies for specific genes. A significance level (p-value) of 0.05 guides our findings. Our analysis uncovers intriguing insights, highlighting specific genetic variations that cause significant changes in binding energy. These variations can greatly affect crucial protein-ligand interactions, influencing the function of these proteins. For instance, variants like THOC5 V579I, NPC1 R1266Q, NPC1 M642I, NPC1 I858V, NPC1 H215R, ABCA1 V825I, ABCA1 R219K, and ABCA1 K1587R display noticeable differences in binding energies. In contrast, variants like THOC5 V525I, MECR F96L, ENPP2 S493P, and CBR4 L70M show minimal changes, implying less impact on binding energy hence protein-ligand interactions. Our study advances our understanding of how genetic variations impact dynamic protein-ligand interactions in terms of binding affinity. These variations could affect protein function, influencing lipid metabolism and heart health. Notably, we identify genetic variants associated with significant changes in binding energies, potentially altering protein-ligand interactions linked to lipid metabolism. These findings extend to lipid balance and the Reverse

Cholesterol Transport pathway, both essential for heart health.The study's importance lies in its innovative approach to studying how genetic variants impact lipid metabolism at the molecular level. By combining virtual screening and statistical analysis, we identify genetic variations that potentially affect protein binding energies, offering insights into treating lipid-related disorders and CVD.

Contents

List of Figures

Chapter 1

Introduction

]

Cardiovascular disease (CVD) is a major global health concern, leading to significant illness and fatalities worldwide. It encompasses a range of conditions, including cardiac arrest, stroke, and arterial disease, contributing to approximately 17 million deaths annually, according to estimates by the World Health Organization (WHO). The rising prevalence of CVD can be attributed to an expanding human population, sedentary lifestyles, unhealthy diets, and an increasing incidence of associated risk factors like high blood pressure, diabetes, and obesity. Certain non-modifiable factors, such as age, race, hereditary predisposition, and family history of CVD, also play a role in influencing individual susceptibility to the disease [1].

Age is a significant risk factor, as the prevalence of CVD tends to rise with advancing age. While women have traditionally been considered at lower risk than males, postmenopausal women become more vulnerable due to hormonal fluctuations. Genetic variables, such as specific gene variations and a family history of CVD, can also impact an individual's risk of developing the condition. The interactions among these risk factors and the complex pathophysiological mechanisms underlying CVD further contribute to its complexity. CVD's consequences can be severe and varied, heavily influenced by genetic factors and epigenetics. As a result, it carries a high fatality rate and significantly impacts individuals' quality of life, leading to debilitating symptoms like fatigue, breathing difficulties, and chest tightness [2]. Several lifestyle factors play a crucial role in the development of CVD, such as unhealthy dietary choices, physical inactivity, smoking, being overweight or obese, having dyslipidemia (abnormal lipid levels), and insulin resistance. Additionally, excessive alcohol consumption poses a significant risk to heart health. Hypertension, elevated glucose levels, and abnormal lipid

profiles are intermediate indicators that, when identified in primary care settings, can signal an increased risk of developing coronary artery disease, stroke, and heart failure [3].

Despite significant advancements in diagnostic methods and therapeutic interventions for CVD, the disease remains a formidable global health threat. While treatments like statins and anticoagulants have improved outcomes, novel and tailored approaches are still necessary to address CVD's complex and diverse nature. Identifying innovative prevention strategies and designing personalized treatment plans are essential to enhance patient outcomes and alleviate the burden on healthcare facilities [4].

The pathophysiology of cardiovascular disease (CVD) is characterized by dysregulated lipid metabolism, leading to increased levels of low-density lipoprotein cholesterol (LDL-C) and reduced levels of high-density lipoprotein cholesterol (HDL-C). This imbalance in lipid levels contributes to the formation of fibrous deposits, composed of collagen, calcium, smooth muscle cells, and foamy cells, within the arteries. Over time, these deposits narrow the arterial lumen and obstruct blood flow. Inflammation also plays a critical role in the pathogenesis of CVD. Inflammatory conditions impair endothelial function, promote plaque formation, destabilize deposits, and increase the risk of thromboembolism—a process where blood clots form and break free, causing blockages in blood vessels. Thrombosis events can further restrict blood circulation, leading to partial or complete blockage of blood vessels. Cardiac failure can result from circulatory remodelling, stiffness, chronic ischemia (insufficient blood supply to the heart muscle), and myocardial rupture [5].

To effectively tackle this significant healthcare burden, developing specific strategies based on a comprehensive understanding of the intricate pathophysiological mechanisms underlying CVD is crucial. Advancing our knowledge of lipid oxidation and the reverse cholesterol pathway holds immense promise in addressing the complexity of heart disease pathology and designing targeted therapies. Developing focused treatment options for CVD requires overcoming the challenges of patient segmentation based on disease vulnerability and gaining insights into the condition's physiology. By identifying individualized risk factors and understanding the molecular processes, personalized approaches can be developed to prevent, diagnose, and treat CVD effectively [6]. An individual's genetic makeup can significantly influence how their body responds to risk factors such as smoking, elevated cholesterol levels, and obesity, impacting their clinical trajectory in cardiovascular disease (CVD). By understanding individuals' genetic predispositions to disease, medical treatment strategies can be developed to identify those in the early stages of illness and provide personalized treatments that are most

effective and have fewer adverse side effects. Lipid metabolism is critical in maintaining the body's overall health and functionality. However, lipid movement and cholesterol balance disturbances can occur in various medical conditions, including metabolic disorders, weight gain, and cardiovascular diseases. Lipid metabolism encompasses the complex processes of lipid production, transportation, retention, and utilization within the body. It is essential for maintaining structural integrity, energy balance, and overall metabolic well-being [7].

Maintaining a delicate balance in lipid uptake and breakdown is crucial, particularly in the small intestine, where these processes primarily occur. The liver plays a predominant role in developing lipids through various enzymatic processes. Acetyl-CoA, derived from the breakdown of carbohydrates, proteins, or fatty acids, is a crucial component in lipid synthesis. The fatty acid synthesis pathway transforms Acetyl-CoA into long-chain fatty acids, which are then used to create essential lipids such as cholesterol, phospholipids, and triglycerides [8]. Conversely, when the body requires an energy source, stored triglycerides are broken down into fatty acids and glycerol in a process known as lipolysis or lipid breakdown. Hormones like glucagon and adrenaline control lipolysis, promoting the study of triglycerides in adipose tissue to release fatty acids [4].

These fatty acids can be utilized as a source of energy by various tissues, including the liver and muscles, contributing to overall lipid balance. To ensure precise control over lipid metabolism, key processes like lipid synthesis, storage, and breakdown are regulated by hormones like insulin and leptin. While hormones such as glucagon and adrenaline stimulate lipolysis to increase the availability of fatty acids for energy production, insulin encourages the storage of lipids. Maintaining optimal cholesterol levels, known as cholesterol homeostasis, is crucial for various physiological activities. Cholesterol, a specific type of lipid, plays essential roles in producing steroid hormones, forming cell membranes, and generating bile acids [9]. The body strictly controls cholesterol levels to maintain cholesterol homeostasis, which is necessary for typical cellular operations. However, an excessive buildup of cholesterol can be dangerous and cause several illnesses, including cardiovascular disease. The body uses a sophisticated set of feedback systems to control cholesterol levels. The equilibrium between cholesterol synthesis, absorption, utilization, and excretion is known as cholesterol homeostasis [10].

A complex interplay of numerous cellular activities, including cholesterol uptake, production, storage, and efflux, regulates cholesterol homeostasis. Both endogenous production and dietary sources provide cells with cholesterol. The enzyme HMG-CoA reductase catalyzes the rate-limiting step in the manufacture of cholesterol, mainly

found in the liver. Harmful feedback mechanisms strictly control the synthesis of cholesterol. The enzyme HMG-CoA reductase is activated when cholesterol levels are low, increasing the production of cholesterol. In contrast, the enzyme is downregulated in high cholesterol levels, which lowers cholesterol production [6]. Very low-density lipoprotein (VLDL) particles, produced in the liver, transport triglycerides and cholesterol to peripheral organs. Through lipoprotein lipase (LPL) activity, VLDL particles become low-density lipoprotein (LDL) particles, commonly known as "bad cholesterol", due to their association with increased cardiovascular disease risk. Cells take up LDL particles via the LDL receptor pathway to maintain cholesterol homeostasis. On the other hand, high-density lipoprotein (HDL) particles, known as "good cholesterol," remove excess cholesterol from cells through the reverse cholesterol transport (RCT) pathway, carrying it back to the liver for excretion [10].

Disturbances in cholesterol homeostasis can lead to dyslipidemia, characterized by abnormal cholesterol levels, and increase the risk of cardiovascular disease. Medications like statins may be prescribed in high-risk individuals to lower LDL cholesterol levels and maintain cholesterol balance. The low-density lipoprotein receptor (LDLR) and sterol regulatory element-binding protein (SREBP) pathway are critical in controlling cholesterol homeostasis and regulating cholesterol synthesis, uptake, and removal [5]. Cardiovascular disease is directly linked to dysregulated lipid metabolism, characterized by irregular lipid levels and cholesterol homeostasis. Elevated low-density lipoprotein cholesterol (LDL-C) is a well-known risk factor for atherosclerosis, the underlying disease of most cardiovascular diseases. LDL-C can infiltrate the artery wall, triggering inflammation and the formation of lipid-rich plaques. Triglycerides, high-density lipoprotein cholesterol (HDL-C), and LDL-C are crucial lipid markers in cardiovascular disease. Low HDL-C and high triglyceride levels are associated with an increased risk of cardiovascular disease, although the precise mechanisms are still under investigation [11].

There comes the reverse cholesterol transport (RCT) pathway, which is a vital mechanism for removing excess cholesterol from peripheral tissues and transporting it back to the liver for excretion. The RCT pathway involves several steps, including cholesterol efflux from cells, particularly macrophages in artery walls. Transporters like ATP-binding cassette transporter A1 (ABCA1) and scavenger receptor class B type I (SR-BI) mediate cholesterol efflux, enabling cholesterol to move from cell membranes to external acceptors, mainly HDL particles. HDL particles, often called "good cholesterol," absorb the effluxed cholesterol. Apolipoprotein A1 (apoA1) is the primary protein in HDL and interacts with ABCA1 to facilitate cholesterol efflux from cells. As HDL particles

absorb cholesterol, they change and convert cellular free cholesterol into cholesterol esters, which are more hydrophobic and can be transported within the core of HDL particles, aided by enzymes like lecithin-cholesterol acyltransferase (LCAT) [12]. After absorbing cholesterol, HDL particles can be delivered to the liver through various mechanisms. One method involves the scavenger receptor class B type I (SR-BI), which directly absorbs HDL cholesterol by hepatocytes, facilitating the liver's uptake of specific amounts of cholesterol without internalizing the entire HDL particle. Another mechanism involves the cholesteryl ester transfer protein (CETP), which transfers cholesterol from HDL particles to triglyceride-rich lipoproteins like very-low-density lipoproteins (VLDL) and intermediate-density lipoproteins (IDL). The transferred cholesterol can be absorbed by the liver or transported to other tissues for utilization or elimination [8].

Once cholesterol reaches the liver, it is processed for excretion. The liver converts cholesterol into bile acids, essential for digesting and absorbing dietary fats. Bile containing bile acids is released into the small intestine, where some bile acids are excreted in faeces, removing cholesterol from the body, while others are reabsorbed in the intestine and enter enterohepatic circulation. Therefore, RCT is considered a preventive measure against atherosclerosis, characterized by the accumulation of cholesterol-rich plaques in arteries. However, the success of the RCT pathway can be influenced by factors like genetics, lifestyle choices, and specific medical conditions [13].

In cardiovascular health, gene expression significantly determines the levels and activity of proteins involved in lipid metabolism and cholesterol homeostasis. Mutations in genes related to lipid metabolism can lead to dysregulation of cholesterol levels, potentially increasing the risk of cardiovascular diseases like atherosclerosis. Conversely, beneficial modifications in these genes may enhance the efficiency of cholesterol removal from arterial walls, promoting cardiovascular well-being. However, it is essential to recognize that the impact of gene expression on protein structure and function is highly context-dependent. Environmental factors, lifestyle choices, and epigenetic modifications can all influence how genes are expressed and how the function of the resulting proteins in specific cellular environments [14].

The mutation is a fundamental process in the theory of evolution, shaping the genetic diversity and origins of life on Earth. In the context of your study on cardiovascular health and lipid metabolism, understanding the role of mutation becomes even more crucial. Genetic variations, such as single nucleotide polymorphisms (SNPs) and structural features like Alu insertions, can significantly affect an individual's susceptibility to cardiovascular diseases. In lipid metabolism, certain genetic variations might

lead to dysregulation of cholesterol and lipid levels, contributing to atherosclerosis and other cardiovascular conditions. These mutations can alter protein structure and function, impacting lipid metabolism pathways and cholesterol homeostasis [15].

Moreover, studying the interplay between gene expression, mutations, and lipid metabolism can lead to advancements in personalized medicine. Researchers can develop targeted interventions and treatment strategies to improve cardiovascular health outcomes by identifying genetic factors influencing lipid profiles. Additionally, this research can shed light on the complex interactions between genetics, lifestyle choices, and environmental factors that influence cardiovascular disease development and progression. Most protein mutations have a neutral impact on their overall behaviour, meaning they do not significantly alter the protein's function. However, modifications with higher similarity between the affected amino acids are more likely to be retained and drive evolutionary changes. It's essential to understand that mutations often have pleiotropic effects, where a mutation positively affecting a protein's activity may simultaneously destabilize the protein, decreasing its soluble and functional form. Additionally, mutations that confer advantages for one function may have adverse or neutral effects on the protein's current function [3].

When a Single Nucleotide Variant (SNV) occurs in a gene and is functionally significant, it is more likely to have a negative effect. Such mutations can disrupt the gene's regulatory mechanisms, impacting its expression, or lead to structural changes at the protein level, affecting its overall function. It is worth noting that only non-synonymous mutations in DNA, which result in amino acid replacements, can bring about these functional changes in proteins. Furthermore, the evolutionary pace of proteins is generally slower than that of DNA changes, suggesting that proteins tend to retain their functional characteristics over longer periods of time. From the literature, it was found that genetic variations in genes involved in the reverse cholesterol transport (RCT) pathway, such as ABCA1 and CETP, can influence HDL metabolism and cholesterol export. The RCT pathway plays a crucial role in protecting against the development of atherosclerosis, also known as cholesterol efflux. This process involves transporting excess cholesterol from peripheral tissues, like macrophages in the artery wall, back to the liver for excretion from the body. Key players in this efflux process are ATP-binding cassette transporters, particularly ABCA1 and ABCG1, which are essential for mobilizing cholesterol and synthesizing HDL [14].

Impaired reverse cholesterol transport and reduced cholesterol efflux capacity have been associated with an increased risk of developing atherosclerosis and cardiovascu-

lar disease. When ABCA1 and ABCG1 expression is dysfunctional or diminished, cholesterol accumulation occurs inside macrophages, forming foam cells, characteristic of early atherosclerosis lesions. Furthermore, disrupted reverse cholesterol transport can disturb cholesterol homeostasis and accelerate the progression of atherosclerosis. In short, the genetic differences in RCT pathway genes have significant implications for HDL metabolism and cholesterol handling, influencing the risk of atherosclerosis and cardiovascular disease. Understanding the role of ABCA1, CETP, and other RCT-related genes in cholesterol efflux can provide valuable insights for developing targeted therapies to improve cardiovascular health and prevent atherosclerotic conditions [4].

Genes such as ENPP2, MECR, THOCS, CBR4, NPC1, and ADRB1 have significant implications for lipid metabolism. ENPP2, a key player in extracellular nucleotide metabolism, converts lysophosphatidylcholine (LPC) to lysophosphatidic acid (LPA), a bioactive lipid that regulates lipid levels, inflammation, and cell division. MECR encodes a trans-2-enoyl-CoA reductase, crucial for mitochondrial fatty acid production, impacting lipid metabolism through the conversion of trans-2-enoyl-CoA to acyl-CoA. THOCS, initially known for its role in RNA processing, has emerged as a potential regulator of lipid-related gene expression, influencing lipid metabolism and homeostasis [16]. CBR4's significance lies in its involvement in lipid aldehyde metabolism, affecting the synthesis of lipid mediators like prostaglandins and leukotrienes. These lipid mediators have profound implications for lipid metabolism regulation and inflammation control. NPC1, a pivotal gene in intracellular cholesterol trafficking, is fundamental in maintaining cellular cholesterol balance. Disruptions in NPC1 function can lead to lipid buildup in late endosomes and lysosomes, culminating in Niemann-Pick type C disease, a severe lipid storage disorder [4]. Finally, ADRB1, encoding the beta-1 adrenergic receptor, takes part in lipolysis activation when adrenergic agonists are present. This activation promotes the conversion of stored triglycerides into free fatty acids, enhancing lipid mobilization and supporting energy expenditure. The intricate involvement of these genes in lipid metabolism underscores their crucial role in maintaining lipid homeostasis and highlights their potential as therapeutic targets for addressing metabolic diseases and disorders [17].

In order to examine lipid metabolism and the reverse cholesterol route in the context of cardiovascular disease, this study proposes an integrated strategy combining gene expression analysis, variant analysed models from previous studies, structural biology, and ligand-based virtual screening. The intersection analysis method utilized in this study has multiple benefits. Firstly, by incorporating prior research and literature on lipid-related pathways and CVD, we used an internal lipid metabolism gene set to

prioritize genes with differential expression and functional significance. This approach helped identify key players in the pathophysiology of CVD and lipid metabolism diseases. Secondly, the intersection analysis approach offers a targeted inquiry by focusing on genes with similar trends across multiple investigations. This enhances the robustness of the findings and reduces limitations from individual studies.

Furthermore, the intersection analysis allows us to identify genes with subtle yet biologically relevant changes in expression rather than solely relying on statistical significance. This is crucial in complex disorders like CVD, where small changes in multiple genes can collectively contribute to the disease's progression. It offers insights into the molecular pathways underlying lipid metabolism and CVD, paving the way for future studies to explore their biological functions and potential as targets for novel therapeutic approaches in treating lipid- and CVD-related diseases. In an attempt to find novel therapeutic targets and ligands for additional experimental validation by methodically filtering and assessing gene expression data, modelling variations, and reducing the compound dataset. This study's research may help create innovative treatments that target dysregulated lipid metabolism in cardiovascular disease and open the door to individualized treatment plans. Cardiovascular disorders, particularly atherosclerosis, have long been recognized as having a genetic component, but obtaining comprehensive information on the specific causal genes has been challenging. The advent of genomic technology and procedures has made it more accessible to identify genes contributing to disease susceptibility and progression. One widely used strategy in candidate gene discovery in gene expression studies using microarrays, which provide valuable insights into disease biology and serve as the foundation for diagnostic tools and treatment plans to significantly enhance human health [1].

Gene expression plays a crucial role in the onset and progression of cardiovascular disease (CVD) by regulating diverse cellular functions involved in cardiovascular health and illness. Dysregulation of gene expression can profoundly impact inflammation-related pathways, lipid metabolism, blood vessel development, and cardiac function. Altered expression of these genes may lead to cellular and molecular changes that contribute to the progression of CVD, mainly through the deregulation of lipid metabolism-related genes [18].

Maintaining cardiovascular homeostasis requires proper gene expression in cell signalling pathways such as the renin-angiotensin-aldosterone system (RAAS) and endothelin signalling. Aberrant gene expression in these pathways can disrupt the balance between vasoconstriction and vasodilation, resulting in hypertension and compromised heart function. Genes controlling the reverse cholesterol transport (RCT) pathway also

play a critical role in lipid synthesis, transportation, and metabolism, especially cholesterol. Adequate gene expression is essential for the effective operation of the RCT pathway, ensuring the production and proper functioning of high-density lipoprotein (HDL) indispensable particles in RCT [14][19].

Furthermore, gene expression influences ATP-binding cassette (ABC) transporters, including ABCA1 and ABCG1, crucial for cholesterol efflux from cells, making cholesterol transfer from peripheral tissues to HDL particles more efficient. Dysregulated expression of these transporters can hinder cholesterol export, leading to cholesterol buildup and atherosclerosis. Additionally, genes involved in lipid metabolism and cholesterol production, such as HMG-CoA reductase and sterol regulatory element-binding proteins (SREBPs), affect the availability of cholesterol for RCT. Dysregulation of these genes can disrupt cholesterol homeostasis and compromise RCT efficiency [12][20].

Moreover, gene expression and RCT are interconnected, as genes implicated in RCT can modulate their activity and influence CVD risk due to genetic variations, epigenetic changes, and environmental influences. Understanding the role of gene expression in RCT is crucial for deciphering the molecular mechanisms of CVD and identifying viable treatment targets. Analyzing gene expression profiles linked to RCT can shed light on regulatory networks and molecular pathways involved in cholesterol metabolism and transport, facilitating the development of novel therapies targeting gene expression or associated pathways to enhance RCT efficacy and reduce CVD risk [17].

Genetic variations, such as single nucleotide polymorphisms (SNPs), can directly impact gene expression levels in lipid metabolism, cholesterol transport, and inflammation, affecting CVD-related complex molecular pathways. These variations can affect the regulatory or coding regions of genes, altering gene expression levels or the structure and function of proteins. Dysregulation of crucial lipid metabolism and RCT proteins can arise in response to genetic variations, leading to lipid homeostasis disturbances and atherosclerosis development [14].

Integrating genetic data with lipid profiles and clinical information can facilitate the development of targeted medications to optimize lipid metabolism and enhance RCT efficiency, thereby reducing the risk of CVD. This approach can also aid in risk assessment, personalized interventions, and risk reduction strategies.

Virtual screening methodologies are valuable tools in drug development for CVD, allowing researchers to explore extensive chemical space and identify potential lead compounds. Virtual screening can help identify substances with desired pharmacolog-

ical properties, such as those targeting enzymes, receptors, or transporters involved in lipid metabolism. Furthermore, virtual screening can aid in developing new medications targeting specific molecular pathways associated with CVD, including endothelial dysfunction, oxidative stress, and inflammation. While virtual screening expedites the early phases of drug development, experimental validation through in vitro and in vivo, tests are crucial to confirm the activity and selectivity of identified compounds [6][21].

Integrating gene expression analysis, variant modelling, structural biology, and ligand-based virtual screening can provide insights into lipid metabolism and the reverse cholesterol route in the context of cardiovascular disease. This study's results focus on identifying novel therapeutic targets and ligands, paving the way for further experimental validation and developing personalized treatment plans targeting dysregulated lipid metabolism in cardiovascular disease.

1.1 Thesis organization

chapter 2 Methodology - This chapter outlines the comprehensive approach used in the study, including data collection and processing methods. It explains how gene expression data were obtained and analyzed and details the modelling of proteins and selection of compounds for virtual screening.
chapter 3 Results and Discussion - This chapter presents the study's findings, including gene expression data analysis, identification of lipid metabolism and CVD-related gene variations, and protein modelling and molecular docking outcomes. The significance of the results and their implications for the research topic is discussed.
chapter 4 Conclusion and Future Directions - The final chapter summarizes the key discoveries, highlighting their relevance for understanding CVD and potential implications for personalized therapeutics. Limitations of the study are addressed, and recommendations for future research are provided, underscoring the study's significance for CVD investigation and patient treatment.

Chapter 2

MATERIALS & METHODOLOGY

2.1 Introduction

The objective of the methods used in this study was to thoroughly analyse lipid metabolism and pinpoint possible therapeutic targets in the context of reverse cholesterol transport (RCT). The critical process known as RCT is essential for preserving lipid homeostasis and reducing the buildup of cholesterol in tissues, which can result in cardiovascular illnesses. To accomplish our goals, a comprehensive strategy combining gene expression data analysis and virtual screening methods was used. In the first part of our research, we used high-throughput gene expression data to identify essential genes and pathways linked to lipid metabolism from relevant biological samples or databases. We sought to understand the molecular connections and regulatory mechanisms controlling lipid metabolism and RCT by examining the expression patterns of these genes. This work uses variant data analyzed in addition to gene expression data analysis and virtual screening approaches to investigate potential genetic factors impacting lipid metabolism and find pertinent pharmaceutical targets for reverse cholesterol transport (RCT). The work intends to advance our knowledge of the intricate interactions between genetics, lipid metabolism, and future therapeutic interventions by integrating variant data to acquire insights into the impact of genetic polymorphisms on lipid-related genes and their association with RCT [10].

We wanted to prioritize targets and find drugs that could modify lipid metabolism and improve RCT effectiveness by utilizing computational approaches and examining gene expression patterns. The study's findings will significantly impact the field of lipid-related cardiovascular illnesses. This research may aid in creating novel therapeutic interventions to enhance RCT and lower the risk of cardiovascular diseases linked to dysregulated lipid metabolism by elucidating the molecular mechanisms underpinning

lipid metabolism and identifying potential drug targets. The combined technique of gene expression data analysis and virtual screening provides a thorough understanding of lipid metabolism, opening the door for creating novel medicines and approaches to deal with lipid-related illnesses.

2.2 Computational Requirements

By successfully addressing the crucial computational needs, the study objectives could be achieved and the successful execution of the necessary analyses, data processing, and development of valuable findings within the allotted timeframe. Completing our investigation was made possible by the successful resolution of these computational requirements.

2.2.1 Linux Operating System Setup

The open-source, Unix-based Linux operating system was used at IIIT Delhi to process the data for this inquiry. Linux was chosen because it is dependable, adaptable, and resilient for handling data-intensive activities. It is the perfect option for data processing and analysis due to its robust command-line interface and comprehensive support for scientific computing. The Linux operating system offered by IIIT Delhi's RayLab met all the requirements for this study. Operating system compatibility was carefully considered during the design and development of the data processing scripts for this project, with a focus on supporting the Windows platform. Although the ensuing data preparation operations were finally carried out on the Linux operating system, careful consideration was paid to the design of these scripts to enable flawless execution and strong performance on Windows platforms. Because of this interoperability, researchers can use the scripts on many operating systems, increasing accessibility and flexibility. The workstations were initially configured by installing and configuring the Linux operating system using the available resources to set up the Linux environment. Administrative access to the computer system was required to facilitate a seamless setup procedure. For researchers, the use of the Linux operating system was essential in enabling the completion of data processing tasks with a high level of dependability and security.

2.2.2 Python Environment Setup

Python was picked because of its simplicity and adaptability because it is a commonly used programming language. Python, so named in honour of the British comedy group Monty Python, was first made available in 1991 to develop a fun and approachable programming language. It is a well-liked option for data analysis and scientific computing due to its straightforward and accessible syntax, extensive standard library, and thriving ecosystem of third-party packages. The most recent version of the language, Python 3, was used in this study because of its stable future and compatibility with contemporary libraries. The Python environment was configured on the Linux workstations to guarantee a smooth workflow. Using the command line, Python has to be installed and set up. The system was updated to ensure Python 3 was running at its most recent version. Depending on the setup and specifications of the research project, either the pip package manager or the conda package manager was used to install the essential Python packages and libraries, such as Pandas and NumPy. These packages offer robust tools for data analysis, computing, and manipulation, improving the Python environment's functionality. A reliable and effective platform for performing the study was made possible by the thorough configuration of the Python environment and the Linux operating system. The study scripts might be utilised on other operating systems by including Windows compatibility, supporting more accessibility and allowing researchers to work with their preferred setup. The adoption of Python made data analysis chores easier and offered a flexible and adaptable programming environment because of its simplicity and broad libraries.

2.2.3 Rstudio

As part of our toolkit for data analysis, we used RStudio Server 2022.02.0 Build 443 in addition to Python. A web-based integrated development environment (IDE) created especially for the R programming language is called RStudio Server. It offers a user-friendly interface for writing, running, and debugging R code, as well as for visualising data and producing reports. We used R and its vast ecosystem of libraries for our study by utilising RStudio Server. The IIITD server, accessed at (http://192.168.2.222:8787/), was used to access RStudio Server. Due to its web-based interface, we could easily access RStudio Server remotely using a web browser on any supported device. Multiple researchers might collaborate in this setting to work on R scripts, share ideas, and facilitate group data analysis. We were able to take advantage of each language's advantages for a thorough data analysis strategy by combining the capabilities of Python and R. Python is suited for a variety of pre-processing jobs,

statistical analysis, and machine learning algorithms because of its adaptability and extensive library of resources. On the other hand, R shines in statistical modelling, visualisation, and advanced data analysis methods.

2.2.4 Jupyter notebook

A robust and well-liked web-based application for data analysis, scientific computing, and machine learning activities is Jupyter Notebook. This open-source software allows Users to build interactive documents with live code, mathematics, images, and text. One of Jupyter Notebook's main benefits is its capacity to combine code execution, data visualisation, and documentation in a single environment. As a result, both academics and teachers favour it as a tool. The seamless integration of code execution, data visualisation, and documentation in a single environment is one of the noteworthy features of Jupyter Notebook. It is comparable to having a virtual lab where we can create and run code, visualise data, and document our complete workflow all in one location. Our Jupyter notebooks are saved with the .ipynb extension, which encapsulates our source code, results, and explanations and makes it simple to share and duplicate our work. We must install Jupyter Notebook on our PC before we can use it. The installation procedure is simple. To install Jupyter Notebook, run the pip command; it will take care of all necessary dependencies of the library used. Once set up, Jupyter Notebook may be launched from the computer's terminal and will launch in the user's default web browser, ready for them to begin their data analysis adventure.

2.3 Tools and web server requirements

2.3.1 Castp: Computed Atlas of Surface Topography of proteins

CASTp (Computed Atlas of Surface Topography of Proteins) is a vital bioinformatics tool for analysing protein 3D structures. It calculates solvent-accessible surfaces, identifying pockets and binding sites critical for ligand interactions. Researchers utilize CASTp to understand protein-ligand and protein-protein interactions, aiding drug design and enzyme-substrate studies. Its user-friendly interface allows easy submission of protein data and visualization of results, making it a powerful resource in structural biology research. The comprehensive insights provided by CASTp have significant implications for advancing our understanding of protein function and facilitating the

development of new therapeutic strategies.

2.3.2 Dynamut2

DYNAMUT2 is an advanced web server that leverages the power of both machine learning and physics-based models to accurately predict and analyze the effects of mutations on protein stability and function. By employing this cutting-edge combination of methodologies, DYNAMUT2 has become a valuable tool in protein biology [22]. Users can easily access DYNAMUT2 by submitting a protein sequence and a list of mutations they wish to study. The web server then employs sophisticated algorithms to generate insightful and detailed graphical representations of the predicted impacts of each mutation. Through these visual outputs, researchers can visualize and comprehend how individual mutations influence the stability and functional capacity of the protein. What sets DYNAMUT2 apart is its comprehensive approach to providing relevant information to users. In addition to offering predictions, the web server offers in-depth details about the underlying models and assumptions used in the analysis. This transparency allows researchers to critically assess the reliability of the predictions and make informed decisions in their studies. The ability to forecast the effects of mutations on protein stability and function is of immense significance in understanding the genetic basis of various diseases and designing personalized therapeutic strategies. DYNAMUT2 serves as a powerful asset in the quest to decipher the intricate relationship between protein structure and function and opens new avenues for drug discovery and biotechnological advancements.

2.3.3 GROMACS: Open-Source Molecular Dynamics Software

An open-source programme called GROMACS (Groningen MAchine for Chemical Simulations) is frequently used for molecular dynamics simulations in various scientific disciplines, including chemistry, biology, and physics. Its main objective is to simulate the behaviour of biological and chemical systems, such as proteins, lipids, and nanoparticles.GROMACS enables scientists to simulate the motion of atoms and molecules based on classical mechanics using classical molecular dynamics principles, making it possible to examine various systems with varied accuracy and depth, from straightforward biological systems to intricate chemical designs [23]. Researchers can examine molecules' kinetic and thermodynamic characteristics using GROMACS, analyse their dynamic behaviour, and improve their geometric structures.GROMACS is written in

C and C++, guaranteeing good performance and effectiveness. It is usable by various users because of its compatibility with several operating systems, including Windows, Linux, and macOS. GROMACS is open-source software released under the GNU General Public Licence, enabling users to download and alter the programme to suit their requirements freely.

2.3.4 Protein Homology Modelling with Modeller

Protein homology or comparison modelling is frequently performed using the standalone programme Modeller. The idea behind homology modelling is that proteins' tertiary structures are more stable than their amino acid sequences. Therefore, proteins with different sequences might yet exhibit observable structural similarities. Since it is difficult, expensive, and time-consuming to obtain experimental structures for every protein of interest using techniques like protein NMR and X-ray crystallography, homology modelling offers a valuable substitute for producing structural models. According to research, naturally occurring homologous proteins and proteins with evolutionary ties have comparable protein structures and sequences. Furthermore, compared to what would be predicted based only on sequence conservation, 3D protein architectures are more evolutionary conserved. Homology modelling uses these observations to produce applicable structural models that support the creation of hypotheses on the composition and function of proteins. The sequence alignment's precision and the template structure selection determine how well a homology model performs. The quality of the output model often degrades when the target protein's and the template's sequence identities diminish. Low sequence identity or sequence alignment and template selection mistakes can lead to significant errors or inaccuracies in homology modelling.

2.3.5 AutoDock Vina: Molecular Docking Software

AutoDock Vina is a widely used molecular docking software in molecular modelling. It plays a crucial role in structure-based drug design by predicting the binding conformations and interactions between a target binding site and a ligand. Researchers can gain valuable insights into drug design and fundamental biological processes by assessing binding energy and affinity using sophisticated scoring functions and algorithms. AutoDock Vina's efficient search technique enables the exploration of the ligand's optimal binding posture within the target site. Its incorporation into research facilitates the examination of molecular docking, allowing for the anticipation and analysis of binding

interactions between target molecules. The data from AutoDock Vina aid in rational drug formulation and ligand structure refinement for therapeutic purposes [9].

2.4 Data preprocessing and visualisation

This study's data preprocessing analysis procedure included several processes and software tools to get insights into gene expression patterns associated with lipid and cardiovascular diseases. Each phase of the analysis is thoroughly explained in the sections that follow.

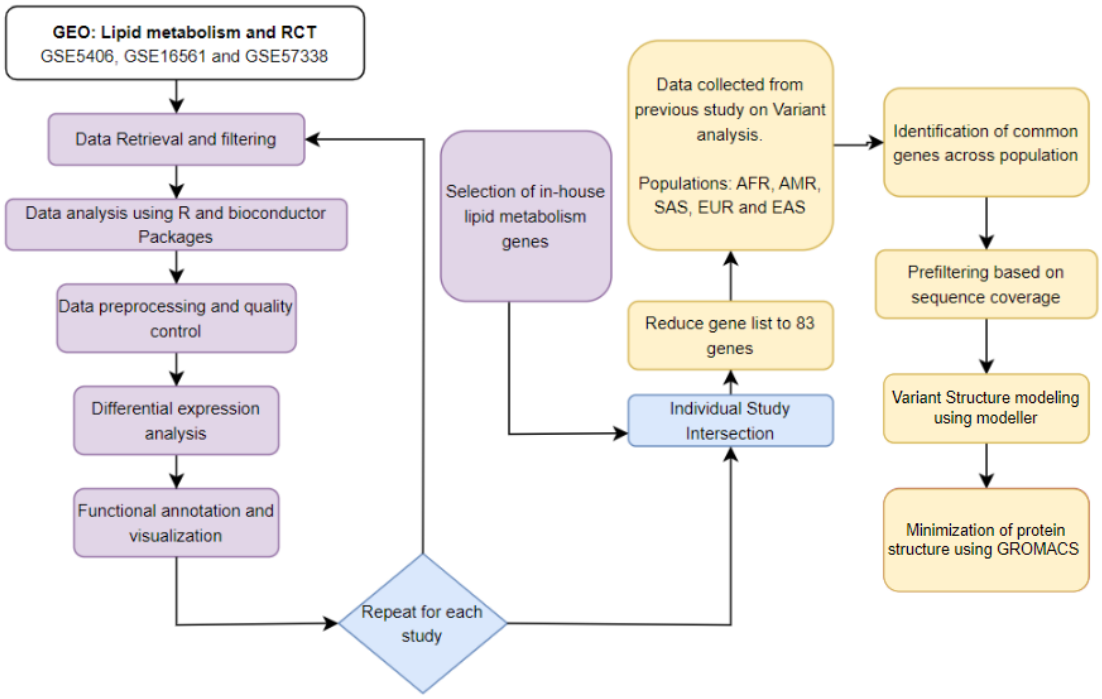

Figure 2.1: Data pre-processing and visualisation workflow

- **Data collection**

  The Gene Expression Omnibus (GEO) database examined the gene expression patterns associated with cardiovascular lipid metabolism [24]. Researchers can access a massive library of publicly available genomic data through GEO to examine

various biological processes and disorders. Multiple cardiovascular diseases were included in a thorough search of gene expression databases. A keyword search was conducted as part of the data collection procedure using specified filters to reduce the results to a proper subset. Datasets with control and case data for comparison analysis were included in this fraction, which used expression profiling by array approaches [25]. We searched the GEO database using the term "cardiovascular + lipid gene expression" to acquire pertinent information for our cardiovascular lipid gene expression study. There were 3822 total results from this search. We ensured that we had access to a wide range of papers about our research issue by searching the GEO database, a frequently utilised repository for gene expression data. We used the GEO website's filtering tools to reduce the number of results. We first choose "Entry type" in the left column as "dataset", And then we were carefully able to concentrate exclusively on datasets with gene expression data. The "Study type" was also set to "Expression profiling by array." This criterion allowed us to focus on studies that used array-based gene expression profiling methods, which aligns with our research objectives. These filters enabled a more manageable dataset for the survey by reducing the number of results to around 150. The comparison of control and case data remained the main focus of further search refinement. They were designed for datasets containing samples from healthy people and those with cardiovascular diseases. This choice aims to shed light on the varied gene expression linked to anomalies in cardiovascular lipids. One can pinpoint particular genes and pathways dysregulated in lipid-related cardiovascular disorders by comparing the gene expression patterns between healthy controls and people with cardiovascular problems. Twenty-seven datasets met these requirements due to this additional filtering step. Finally, we took three datasets that were pertinent to our investigation. Gene expression analysis and differential expression analysis were independently applied to each dataset. The objective was to find the genes whose expression had significantly changed in each study. Samples from people with human ischemic cardiomyopathy, idiopathic cardiomyopathy, and non-failing controls are included in the first dataset, GSE5406 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE5406). This dataset offers valuable information on gene expression concerning certain particular cardiovascular diseases. Ischemic cardiomyopathy is a term used to describe a condition in which the heart muscle weakens due to decreased blood flow and oxygen delivery brought on by coronary artery disease. Idiopathic cardiomyopathy, on the other hand, refers to a condition where the cause of heart On the other hand, idiopathic cardiomyopathy is a condition in which the reason for heart muscle malfunction is not recognised. In this dataset, non-failing controls are those with no de-

tectable cardiac problems. The second dataset, GSE16561, primarily investigates gene expression in peripheral whole blood RNA after an ischemic stroke (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE16561). Studying the changes in gene expression in peripheral blood after an ischemic stroke can shed light on the broader implications and molecular changes brought on by this cardiovascular event. The microarray technology, which allows for the simultaneous assessment of the expression levels of thousands of genes, was used to create the gene expression data in GSE16561. This dataset provides essential knowledge about the molecular changes in response to stroke by analysing gene expression patterns in peripheral blood samples from ischemic stroke patients. Although it has no direct connection to the gene expression of cardiovascular lipids, it provides information about how gene expression varies during a cardiovascular event. The third dataset, GSE57338, uses microarray technology to find new heart failure-related myocardial gene expression profiles (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE57338). Using microarray, a high-throughput sequencing technique that enables the quantification of RNA molecules in a sample, the gene expression data for GSE57338 was produced. Compared to conventional microarray technology offers a more thorough and complete assessment of gene expression. This dataset offers a more profound knowledge of the gene expression alterations linked to the disease by characterising the transcriptome of cardiac tissue samples from patients with heart failure. A sample of pertinent datasets that fit our research goals was found by systematically searching the GEO website and using particular filters. These databases contain insightful information on the molecular mechanisms underlying cardiovascular disorders such as cardiomyopathy, stroke, and heart failure, as well as helpful gene expression data relating to anomalies in cardiovascular lipids.

- **Preprocessing Using R and Bioconductor packages**

  The R programming language and specific Bioconductor packages were used for data analysis. R version 4.2.2 and the following packages were used for the analysis: Biobase 2.58.0, GEOquery 2.66.0, limma 3.54.0, and DESeq2 1.38.3. Utilising the Limma software, the analysis concentrated on differential expression analysis [26].

- **Data Preprocessing and Quality Control**

  The GEOquery software downloaded the gene expression information from the

all dataset. The data was cleaned, normalised, and transformed for sample comparison using the log2 function. In order to remove problematic data points and improve the study's dependability, quality control methods were put in place [27].

- **Differential Expression Analysis**

  The lmFit function from the limma package fits a linear model to the processed gene expression data. To contrast the control and case samples, contrasts of interest were put up. Relevant genes were then identified based on a modified p-value threshold of 0.05 and statistics computed using the Bayes function. Using the top-table function, a table of differentially expressed genes was created. These genes were then further filtered using an adjusted p-value threshold of 0.05 [16].

- **Functional Annotation and Visualisation**

  The hgu133plus2.db software assigned gene symbols to the differentially expressed genes. The plots software created a heatmap visualisation to show the differentially expressed genes' expression patterns. Other visualisations were created to evaluate the statistical significance and distribution of gene expression changes, including histograms, Venn diagrams, Q-Q plots, volcano plots, and MD (mean difference) plots.

- **The Intersection of Differential Expression Results with In-house Lipid Metabolism Genes**

  We used a curated list of internal lipid metabolism genes to look into the potential functions of genes in cardiovascular disease (CVD) and lipid metabolism. Genes closely examined and connected to lipid and reverse cholesterol pathways made up this group. These genes were chosen based on previously published research that showed they were involved in lipid metabolism and CVD.

In this study, three separate investigations were concluded, each focused on analyzing the relationship between gene expression and lipid metabolism. For each investigation, we independently performed an intersection analysis to identify genes that displayed significant differential expression and were linked to lipid pathways. The reason for conducting intersection analyses in each investigation was to identify gene candidates consistently associated with lipid metabolism and cardiovascular diseases (CVD). By examining the overlapping genes across the

three analyses, we sought to establish more reliable and consistent results, reducing the likelihood of spurious findings and enhancing the confidence in the potential gene candidates related to lipid metabolism and CVD.

- **Consolidation of Intersection Results**

  We pooled the lists of intersected genes received from each investigation and removed duplicate genes to synthesise the results from the three studies. With the help of this procedure, we came up with a final selection of 83 distinct genes that consistently displayed differential expression and associations with lipid and reverse cholesterol pathways. A carefully chosen group of candidates with possible contributions to CVD and lipid metabolism comprised these 83 genes.

- **Identification of Common Genes in Cardiovascular Disease (CVD) and Lipid Metabolism-Genetic Analysis across Populations**

  We thoroughly investigated the 83 common genes discovered in the previous step to clarify the genetic influences on cardiovascular disease (CVD) and lipid metabolism across various populations. Our goal was to shed insight into population-specific aspects of CVD and lipid metabolism by examining these genes' presence and potential role in various ethnic groups.

- **Intersection Analysis of Variant Data**

  We used variation data from six different populations, including African (AFR), European (EUR), South Asian (SAS), East Asian (EAS), and Admixed American (AMR), to achieve our goal. With a focus on the reverse cholesterol pathway linked to cardiovascular disease, the analysis covered 182 genes involved in lipid metabolism. The 1kG and IndiGen sample datasets were used to find 7,025 non-synonymous SNPs in the 182 genes and 2,364 SNPs in 58 genes associated with the reverse cholesterol pathway. These SNPs were further filtered based on allele frequency to get a final set of variants for each population. Amino acid substitution frequencies and patterns were examined across several races to comprehend the effects of these polymorphisms. With variations in prevalence and frequency across populations, the study identified diverse patterns of amino acid exchanges. Exchanges unique to specific communities suggest that those populations have different evolutionary histories and genetic ancestries. Additionally, mutability studies revealed variances in populations' rates of mutation. The study

also looked at how these alterations affected the stability and structure of proteins. The findings suggested that the variations had stabilising and destabilising effects, possibly affecting protein function. The hydrogen bonding network, solvent accessibility, and protein-protein interactions were the subjects of additional investigation to learn more about the functional effects of the discovered variations. The study found modifications to the solvent accessibility, changes in the hydrogen bonding network, and probable disruptions in protein-protein interactions, all pointing to potential functional effects of the variations. The study also showed the intricate relationships between the analysed genes and the pathways involved in lipid metabolism. This study advances the understanding of the role of genetic polymorphisms in cardiovascular disease susceptibility by offering a broader view of the global genetic architecture of lipid homeostasis and may provide methods for preventative interventions [28]. The results highlight the need to consider varied populations and their genetic variety when examining the genetic underpinnings of diseases and disorders associated with lipid homeostasis. By combining the 83 common genes with the variation data of each community, we aimed to find genes that were consistently present and may have functional significance across a range of ethnic backgrounds.

- **Consistent Genes across Populations**

  We discovered a subset of genes through our intersection analysis that showed constant presence in all six populations. The significant degree of conservation shown by these genes suggests that they may modulate CVD and lipid metabolism pathways in people with different genetic backgrounds [28].

- **Refinement of Gene Set based on Reverse Cholesterol Pathway Relevance** Our attention was next directed towards choosing genes that specifically contribute to the reverse cholesterol pathway once we had obtained a refined collection of 83 unique genes from the intersection of variation data with the common genes. To do this, we implemented a selection procedure that used internal expertise. By using this selection method, we were able to isolate a more focused group of 25 genes that displayed potent connections with the reverse cholesterol pathway in all examined populations. These genes stood up as promising candidates for additional research in CVD due to their high possibility of involvement in cholesterol control and metabolism.

- **Prefiltering Based on Sequence Coverage**

  A prefiltering phase based on sequence coverage was applied to further reduce the gene set. To guarantee that the chosen genes have robust and trustworthy sequence data available,the raw sequence data underwent essential filtering steps to ensure data quality and relevance. The filters applied are outlined as follows:

  Accessibility of Protein Crystal Structures:

  - **Only protein crystal structures publicly available and accessible were considered for the analysis.** Protein Crystal Structures with Sequence Coverage 70 %:
  - **To ensure sufficient coverage of the protein structures, only those with a sequence coverage of 70% or more were included.** Non-synonymous Single Nucleotide Variations (SNVs) Observed in Populations with Allele Frequency 10%:
  - **Non-synonymous SNVs, which lead to changes in amino acids, were selected for the analysis.**
  - **SNVs observed in populations with an allele frequency of 10% or higher were considered for further investigation.**

  By employing these rigorous filtering criteria, the data was refined, enhancing the reliability and significance of subsequent structural-level analyses. A final collection of 10 genes that demonstrated relevance to the reverse cholesterol pathway and sufficient sequence coverage for further analysis emerged from this pre-filtering stage.

- **Focus on the Subset of 10 Genes for In-depth Investigation**

  We sought to learn more about the distinct functions, regulatory mechanisms, and potential consequences of this subset of 10 genes in the context of CVD and lipid metabolism by concentrating on them. This subset of 10 genes resulted from the junction of population-specific common genes and the reverse cholesterol pathway. We were able to prioritise the genes most likely to be directly involved in the reverse cholesterol pathway, allowing for a more focused analysis of their functional relevance and potential as CVD treatment targets.

- **Generation of Structural Models**

  We used computational tools like Modeller to create structural models to obtain insight into the structural implications of genetic variants in the chosen genes. These models represented the original protein structures linked to lipid metabolism and cardiovascular disease (CVD). Using known protein structures and sequence alignments, Modeller, a popular software programme, made creating precise comparative protein structure models easier.

- **Integration of Variant Data into Structural Models**

  The last phase entailed incorporating variant information, such as discovered genetic differences from various populations, into the matching native structure models. We included single nucleotide polymorphisms (SNPs) in the structural models of the genes based on the variant data. Because of this integration, We could mimic the structural effects of these genetic variants on the proteins' three-dimensional conformation.

- **Bridging the Gap: Genetic Variations and Structural Consequences**

  We sought to close the gap between genetic differences and their structural effects by incorporating variant data into the native structure models and performing further analysis. This method made it easier to fully comprehend how particular genetic variants might affect the protein structure and function linked to CVD. The underlying mechanisms and prospective therapeutic targets for intervention were fundamental discoveries supplied by it. In conclusion, the creation of structural models, integration of variant data, and structural analysis provided information about the structural effects of genetic differences in the genes linked to lipid metabolism and CVD. We were able to study the possible effects of particular genetic polymorphisms on protein structure and function using this method, opening the door to more research into their functional effects and the creation of specialised treatment plans.

2.4.1 Ligand Selection and Processing

The next step after collecting the native structure and variant model was to gather ligands from various sources associated with cardiovascular disease. We gathered ligands

from databases like CHEMBL(https://www.ebi.ac.uk/chembl/), ZINC12 (for natural products only)(https://zinc12.docking.org/) , IMPPAT (phytochemicals)(https://cb.imsc.res.in/imppat/) , and DrugBank(https://go.drugbank.com/) . Small organic compounds and their related biological activity are the main emphasis of CHEMBL, a comprehensive database of bioactive chemicals. Chemical structures, experimental bioactivity data, drug-like characteristics, and citations to academic works are only a few of the many details it offers. Using this database, scientists can investigate and evaluate chemical substances and their potential therapeutic applications for drug discovery and pharmaceutical research.

On the other hand, the widespread database ZINC focuses on giving users access to commercially accessible chemicals for virtual screening and drug discovery. It provides many tiny organic molecules from suppliers, including natural and artificial substances. Thanks to its extensive library of buyable compounds, ZINC is a valuable tool for virtual screening studies and compound selection in drug development initiatives. The database IMPPAT is devoted to phytochemicals, chemical substances derived from plants. It serves as a comprehensive data database on phytochemicals, including information on their chemical compositions, biological functions, therapeutic potential, and sources in various plant species. IMPPAT makes exploring and examining phytochemicals easier, revealing details about their possible therapeutic uses, health advantages, and contributions to conventional medicine. Last but not least, DrugBank is a vast database that contains thorough data on medications, drug targets, and drug-related data. It includes many drug listings for approved, investigational, and medication candidates. DrugBank offers in-depth details on drug targets, pharmacological effects, modes of action, therapeutic indications, drug metabolism, interactions with other drugs, and clinical data. This database supports drug development, drug repurposing, and the comprehension of the pharmacological properties of existing medications, serving as a valuable resource for researchers, doctors, and drug developers.

With access to chemical compounds, information on bioactivity, and other pertinent data, these databases collectively play critical roles in drug discovery and pharmaceutical research. They are valuable resources for locating prospective drug candidates, comprehending how they interact with biological targets, and investigating the therapeutic potential of natural substances and phytochemicals. A library of roughly 150,000 chemicals in SDF (Structure-Data File) format was produced as a result of this thorough gathering. It was decided to use Padelpy, a programme that computes molecular descriptors, to lessen the complexity of this substantial compound library [29]. Padel Descriptor software was used to gather a wide range of information about the chemical compounds, and 1,875 descriptors, including 1D, 2D, 3D, and fingerprint descriptors,

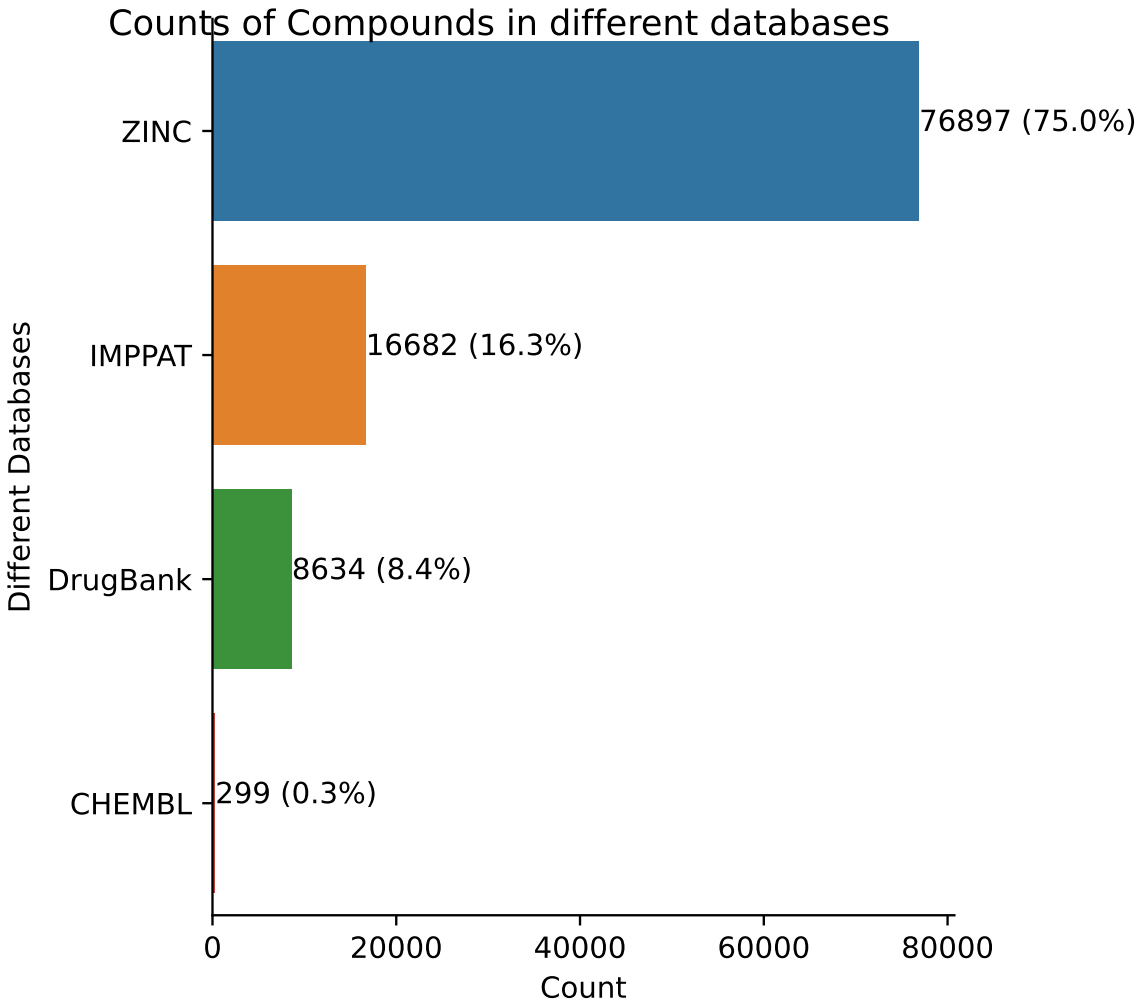

Figure 2.2: Compounds count collected from different database

were extracted for all the chemicals. Each kind of descriptor offers remarkable insights into various properties of the substances. Based on their molecular formulas, basic details on the compounds, such as atom counts and molecular weight, are provided by 1D descriptors. These descriptors make understanding the molecules' composition and connection possible. The compounds' molecular architecture and atom connectivity are the main topics of 2D descriptors. They record details on connectivity patterns, bond kinds, and other topological properties.

These descriptors make characterising the compounds' connectivity and structural

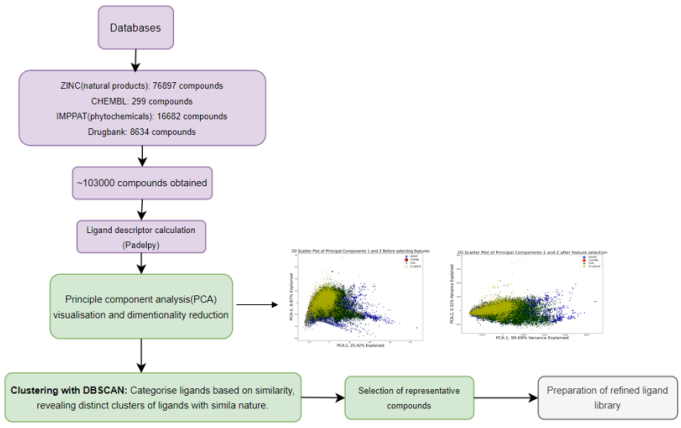

Figure 2.3: Ligand Selection and processing

characteristics possible. The three-dimensional arrangement of the atoms in the mixes is taken into account by 3D descriptors. They shed light on the compounds' conformation, spatial orientation, and structure. These descriptions provide a more thorough understanding of the three-dimensional structure and characteristics of the compounds. On the other hand, fingerprint descriptors are binary descriptors that encode the presence or absence of particular chemical fragments or substructural characteristics inside the compounds. They can distinguish between similar or dissimilar compounds based on shared or distinctive substructures. The features of the combinations can be more fully represented by incorporating all of these descriptor kinds.

Using 1D, 2D, 3D, and fingerprint descriptors and the compounds' composition, structural characteristics, three-dimensional arrangement, and substructural aspects enables a multidimensional understanding of the compounds. The information that can be used for later studies, like virtual screening, similarity searching, or machine learning models, is increased by extracting a wide range of descriptors. It offers a more thorough definition of the substances, making it easier to spot any possible connections, parallels, or contrasts between the chemical structures under inquiry. A CSV (Comma-Separated Values) file format was used to store the outcome of the operation. Principal Component Analysis (PCA) was used to visualise the descriptors and reduce the dimensionality of the data. PCA is a popular multivariate statistical method that projects high-dimensional data onto orthogonal axes to enable visualisation. PCA was used in this work to examine the underlying structure and variability within the dataset using the extracted descriptors as input. Various PCA analysis combinations were used to find the features representing the most variability in the data. The top 10 features

from the first principal component (PC1) and the top five features from the second principal component (PC2) were chosen by examination of the scree plot, eigenvalues, and cumulative explained variance. The chosen features explained approximately 99.71% of the overall variability seen in the dataset. The dataset's dimensionality was successfully decreased by keeping the most valuable features contributing to overall variance. Reducing the number of dimensions made it possible to describe the compound descriptors more succinctly and effectively while keeping the crucial trends and variations in the data.

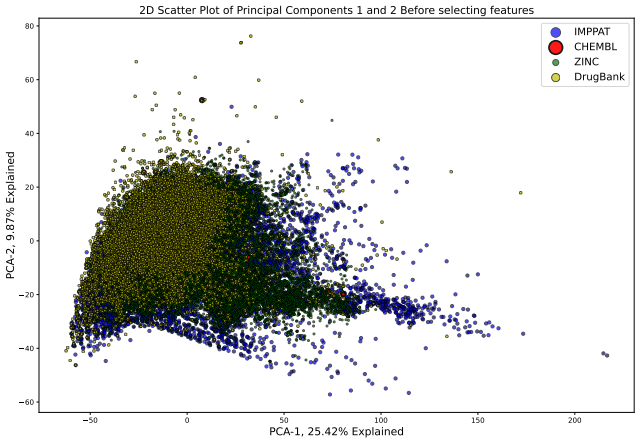

Figure 2.4: PCA Analysis before selecting features

Compressing the original dataset using the chosen principal components and their accompanying features made it possible to visualise the relationships and patterns among the chemicals. PCA made it possible to identify the primary sources of variance within the data, assisting in further studies and more succinctly interpreting the compound descriptors. The following phase involved grouping the compounds according to their descriptors after the PCA analysis and feature selection. Initially, the K-means clustering technique was attempted to be used. However, K-means clustering could have been more effective in successfully dividing the chemicals into meaningful clusters due to the significant data points in the dataset. The Density-Based Spatial Clustering of Applications with Noise (DBSCAN) technique was used as an alternate strategy. Data points are grouped using the density-based clustering algorithm DBSCAN depending on how close they are to one another in the feature space. DBSCAN can accommodate unevenly shaped clusters and noise points, unlike K-means clustering, which has a pre-

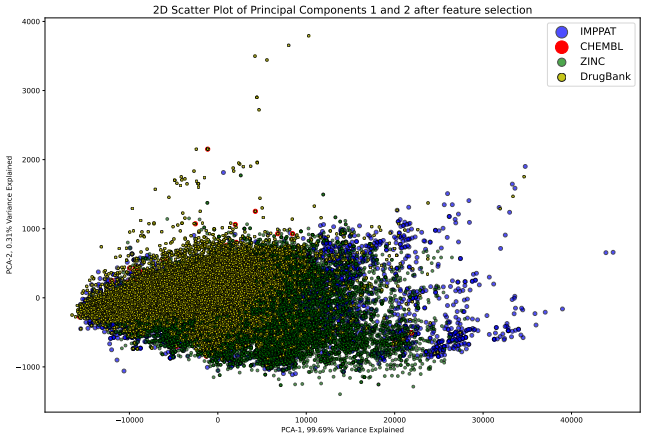

Figure 2.5: PCA Analysis after selecting features

determined number of collections that must be present(Schubert et al. 2017). Based on how similarly the chemicals behaved in the feature space, the compounds were clustered using DBSCAN into almost 9,000 clusters. A further 45,000 chemicals were classified as "noise points," meaning they showed little or no significant relationship to any cluster. DBSCAN clustering made finding cohesive clusters of compounds with shared characteristics possible while considering noise points. This method offered a more adaptable and flexible clustering technique, thinking the complex dataset structure and reflecting the natural variability within the compounds. The DBSCAN clustering results were used as a starting point for additional investigation and study of the compound groups [30]. Researchers can acquire insights into the linkages and similarities among the compounds by locating clusters with shared traits, possibly indicating subgroups or classes of compounds that display particular features or attributes of interest.

After grouping the 9,000 clusters with the DBSCAN technique, representative chemicals from the clusters were sought out. This procedure aimed to choose molecules that displayed traits typical of the different groups. One or more representative compounds were selected from each cluster based on their characteristics and traits. This selection procedure led to creation of a smaller library with about 54,000 chemicals. This collection included the unclustered compounds that did not align with any particular cluster and the representative compounds from the clusters. The final compound library's retention of a wider variety of chemical structures was made possible by the addition of unclustered compounds. The selected sample compounds acted as models for their par-

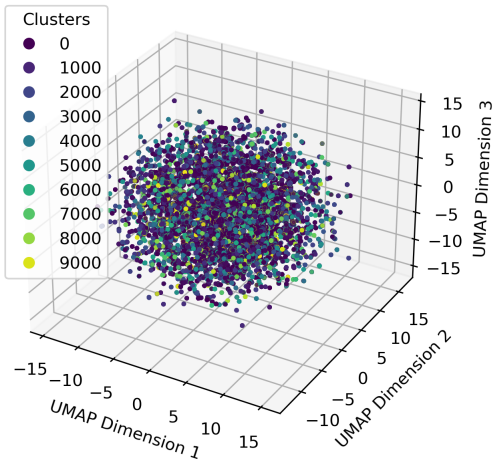

Figure 2.6: UMAP 3D visualisation clusters predicted by DBSCAN algorithm

ticular clusters, capturing the essential traits and attributes of the compounds within each group. The resultant dataset kept the crucial details and diversity while drastically lowering the complexity and size of the original chemical library by including these sample compounds in the reduced library. The smaller library of over 54,000 compounds offers a more manageable selection of compounds for virtual screening and subsequent research, serving as an essential resource for additional analysis and testing. In the context of cardiovascular disease research, these representative compounds have the potential to aid in the discovery of prospective lead compounds or bioactive molecules by providing insights into the rich chemical space investigated throughout the clustering process. Because of this, the methodology for ligand selection and processing included gathering ligands from various databases, reducing the library through descriptor calculations, visualising the data using PCA analysis, clustering the compounds using DBSCAN, choosing representative compounds, and preparing the refined ligand library for virtual screening experiments. This methodical technique is intended to find potential candidates for more research in cardiovascular disease.

2.4.2 Virtual screening and analysis

The first step in the virtual screening method was to gather binding energy data for a particular gene's native and variant structures. Autodock Vina was used to calculate the relevant binding energies of 54,162 molecules. The binding energy differences were then calculated for each compound by deducting the native energy from the variant energy. The 23 proteins, comprising both native and variant models, were virtually screened and docked using Autodock Vina and parallel processing on a server with a smaller dataset of 54,000 compounds. The objective was to forecast the compounds' binding affinities to the protein targets and understand how variants affect the binding energy with ligands.

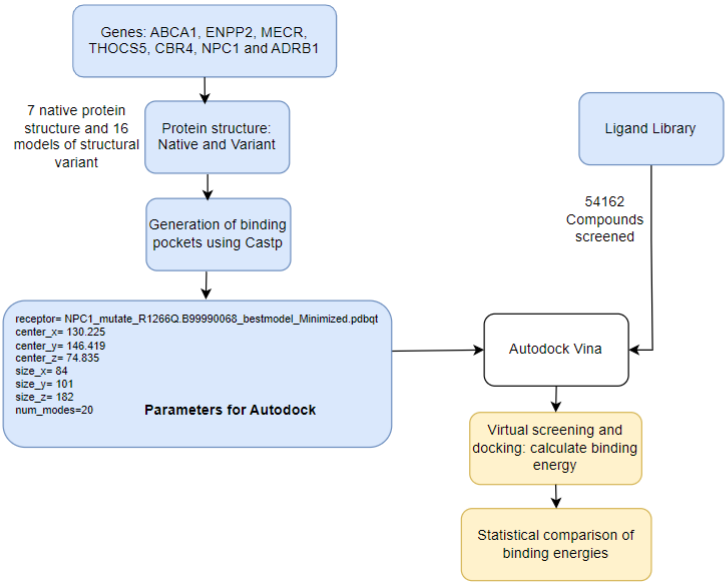

Figure 2.7: Virtual Screening Workflow

A histogram or density map was created to show how the binding energy differences were distributed. This visual depiction made it easier to see which ligands between the variant and native structures had more negative or positive binding energy differences. The plot gave important information on the types and frequency of changes in binding energies by variants. In the virtual screening process, a meticulous statistical analysis of ligand-protein interactions was conducted to assess polar contacts and disruptions between the ligand and protein structures, considering both the original and variant

forms. The primary objective of this investigation was to comprehensively evaluate the impact of genetic variations on ligand binding. Significant modifications induced by the mutations were identified and characterized by comparing these interactions. This rigorous virtual screening approach aimed to determine whether the presence of mutations in the genes under investigation led to any disturbances or alterations in the binding affinity of the ligands. The statistical comparison of binding energies using tools like the Kolmogorov-Smirnov (K-S) test allowed for a robust assessment of the differences in protein-ligand interactions between the wild-type and variant structures. The statistical analysis conducted in this study provides critical insights into how genetic variations can influence ligand binding in the context of cardiovascular disease.

Chapter 3

RESULTS

3.1 Differential Expression Analysis Results in CVD and Lipid Metabolism

3.1.1 Data Preprocessing and Quality Control

The study aimed to analyze transcriptomic differences in Cardiovascular Disease (CVD) by selecting and processing three open-access microarray databases. This approach provided valuable insights into the molecular alterations associated with lipid metabolism pathways and their relevance to CVD. By leveraging microarray technology, we gained a comprehensive understanding of the gene expression patterns involved in cardiovascular pathology, shedding light on potential biomarkers and therapeutic targets for addressing this complex and prevalent disease. In this study, we analyzed gene expression in Cardiovascular Disease (CVD) using three datasets from GEO: GSE5406, GSE16561, and GSE57338. The investigation provided insights into molecular alterations related to lipid metabolism pathways and CVD, aiding in biomarker discovery and potential therapeutic targets. In the context of our scientific research, data preprocessing and quality control are vital steps in gene expression analysis, ensuring the accuracy and reliability of our findings. In our study, we diligently performed robust data preprocessing using R version 4.2.2 and essential R packages such as Biobase 2.58.0, GEOquery 2.66.0, and limma 3.54.0. As a result, the gene expression data from the three datasets (GSE5406, GSE16561, and GSE57338) exhibited a normal distribution in their expression levels and consistent median values across all samples. These observations confirm the suitability of the datasets for further statistical analysis.

The study implemented a rigorous pre-processing pipeline to ensure the accuracy and reliability of the gene expression data. In our study, we chose not to combine

multiple data sets as they were generated from different laboratories using diverse platforms. This decision was made because it is unlikely that these data sets would have the same confounding factors. Instead, we employed specific data preprocessing techniques, such as quantile normalization and log2 transformation to stabilize the variance of gene expression measurements, making the data suitable for statistical analyses and improving the comparability of expression values across genes and samples. A check for log-transformation appropriateness was performed, and genes with zero or negative expression values were handled by assigning them a value of NaN (Not a Number) to avoid distortion of downstream analyses. It allowed us to observe the overall pattern of gene expression and evaluate the success of normalization in achieving a consistent and suitable data distribution.

Implementing rigorous quality control measures provided us with a high level of confidence in the reliability of our study's results. Through careful pre-processing of the raw microarray data, we ensured the integrity and accuracy of the gene expression measurements. The consistency observed in the gene expression distributions, balanced histograms following normalization, and stable mean-variance trends collectively indicated the successful normalization process and highlighted the suitability of the data for further analysis. With these quality control assessments yielding positive outcomes, we could confidently proceed with the differential expression analysis. The identification of statistically significant changes in gene expression between the control and disease groups was bolstered by the robustness of the pre-processing steps. Further, the utilization of the Limma package, with its sophisticated statistical methods and FDR adjustment for multiple testing, further enhanced the reliability of the differential expression results.

3.1.2 Identification of Differentially Expressed Genes in Cardiovascular Disease (CVD)

In this study on Cardiovascular Disease (CVD), we carefully curated sample groups to ensure robust and biologically meaningful analyses. Gene expression data comprised samples from both Control and Disease conditions. A sample annotation vector accurately reflected each sample's disease status, with 1s and 0s representing Disease and Control groups, respectively. This clear distinction allowed downstream analyses to effectively capture gene expression differences associated with CVD. The assignment of group labels enabled the organization of gene expression data based on two distinct conditions: control vs. disease, forming the foundation for subsequent differential expression analysis. The aim was to identify genes significantly altered between the two groups, potentially implicated in CVD pathogenesis. We employed the limma package

in R to fit a linear model to gene expression data while incorporating group membership information. The design matrix modelled gene expression's association with the presence or absence of CVD, effectively controlling for potential confounding factors and obtaining accurate results. Contrasts were established to evaluate gene expression differences between control and CVD groups based on log-fold change, quantifying expression alterations between the conditions. The e-Bayes method calculated adjusted p-values, considering multiple testing and providing statistical significance for each gene. The resulting list of differentially expressed genes, with adjusted p-values below the significance threshold (e.g., 0.05), included potential candidates associated with CVD, involved in relevant biological processes and pathways.

The statistical significance of these differential expression patterns was determined using the e-Bayes method, which applies empirical Bayes smoothing to standard errors of estimated log-fold changes, enhancing analysis robustness, especially with limited sample sizes. Genes with adjusted p-values 0.05 were considered significantly differentially expressed between CVD and control groups. The identified differentially expressed genes serve as valuable resources for further investigations, representing potential candidates pivotal in CVD development and progression. Researchers can perform functional analysis and pathway enrichment studies to gain deeper insights into their roles in CVD. Moreover, these differentially expressed genes are promising CVD biomarkers, indicating potential use in early disease detection, disease progression monitoring, or predicting treatment responses. Furthermore, knowledge gained from this analysis may inform the development of targeted therapeutic strategies addressing CVD's underlying molecular mechanisms.Combining group membership and differential expression analysis using the limma package significantly contributed to our understanding of CVD's molecular landscape.

Implications of Differentially Expressed Genes: Identifying differentially expressed genes in each dataset offered valuable insights into the molecular basis of CVD and lipid metabolism within distinct biological contexts. These genes represent potential candidates for various aspects of research, such as biomarker discovery, therapeutic target prioritization, or investigation of critical regulatory factors governing the pathways of interest [17].

Moreover, the independent analysis of multiple datasets significantly enhanced the robustness of our findings. It allowed us to validate and cross-validate the identified differentially expressed genes across diverse experimental conditions, thereby minimizing dataset-specific biases. As a result, we achieved a more comprehensive understanding of the underlying molecular mechanisms involved in CVD and lipid metabolism-related disorders.

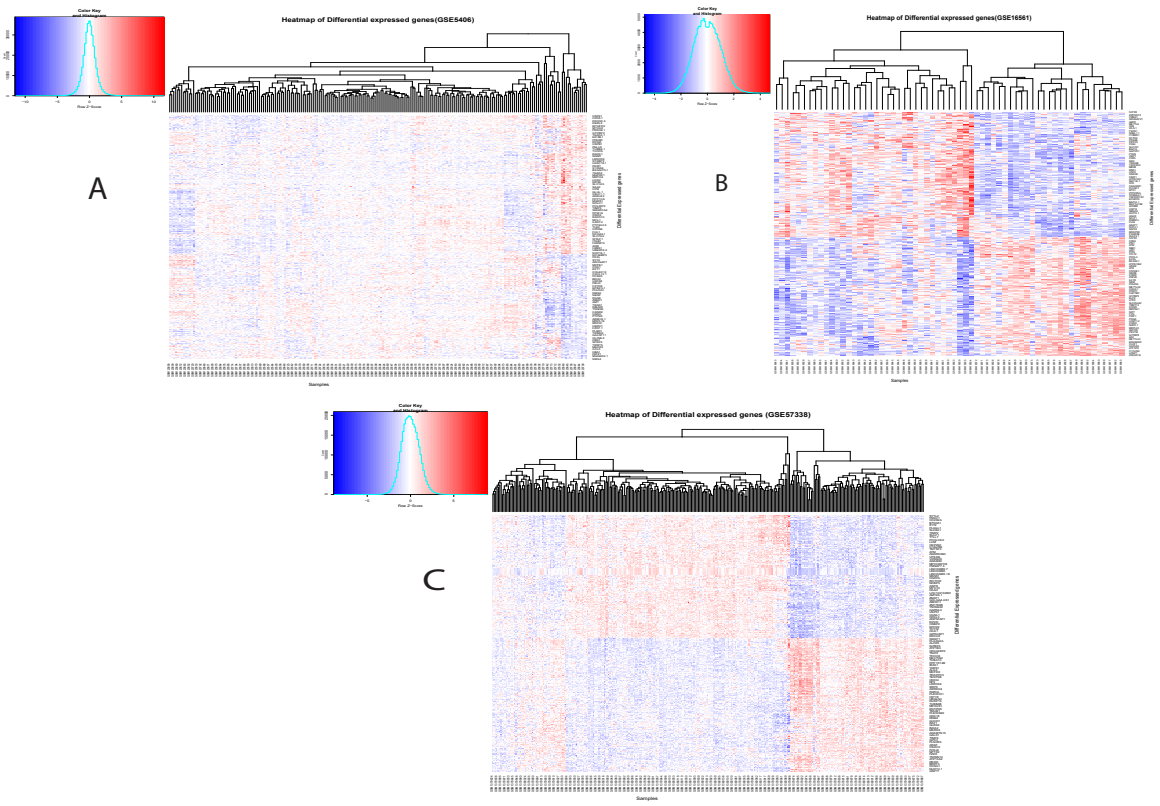

Figure 3.1: Heatmap showing Differential expressed genes A) GSE5406 B)GSE16561 C)GSE57338

Integration of Results: By integrating the results from these three independent studies, we identified common genes and pathways that may play pivotal roles in multiple contexts. This integrative approach holds immense potential for future research and therapeutic interventions, as it unveils shared molecular targets that could have broad implications for developing novel treatments and preventive strategies in the context of CVD and lipid homeostasis. Overall, our differential expression analysis and subsequent integration of results have contributed significantly to our understanding of the molecular landscape associated with Cardiovascular Disease and lipid metabolism. These findings not only provide important insights into the pathogenesis of CVD but also open new avenues for further research, with the potential for the development of targeted therapies and personalized interventions in the fight against cardiovascular

disorders.

Study 1 - GSE5406 (RNA-Seq Identifies Novel Myocardial Gene Expression Signatures of Heart Failure):

In Study 1, we investigated the differential gene expression associated with heart failure using the GSE5406 dataset. Our main objective was to uncover specific gene expression patterns that could provide valuable insights into the molecular mechanisms underlying heart failure. To ensure the quality of the data, we conducted rigorous preprocessing and quality control steps on the raw microarray data, as described earlier. Following data preprocessing, we utilized the limma package, a widely-used statistical method, to identify genes significantly differentially expressed between the heart failure and control groups. The analysis yielded a list of 1691 differentially expressed genes in GSE5406. Among these genes, 720 were upregulated, indicating increased expression levels, while 971 were downregulated, suggesting decreased expression levels. The fold changes of the differentially expressed genes were more significant than —1— (absolute fold change >2), and the significance level was set at a p-value ¡ 0.05. To visualize the overall gene expression patterns, we created a heatmap of the top 500 differentially expressed genes shown in Figure 3.1. The heatmap provided a comprehensive view of the gene expression changes between the heart failure and control groups, highlighting distinct gene clusters associated with the disease state. Additionally, we generated a volcano plot to assess the statistical significance of the differentially expressed genes. The volcano plot illustrated each gene's log-fold changes (effect size) against their corresponding p-values as showin in Figure 3.2 Genes significantly upregulated or downregulated were prominently displayed on the plot, representing potential candidate genes with substantial relevance to heart failure.

Study 2 - GSE16561 (Gene Expression Analysis of Peripheral Whole Blood RNA Following Ischemic Stroke):

Study 2 investigated the gene expression changes in peripheral whole blood RNA following ischemic stroke using the GSE16561 dataset. The primary aim was to identify differentially expressed genes associated with the pathophysiology of ischemic stroke and to gain insights into the peripheral blood gene expression response to this neurological event. Before the analysis, we performed preprocessing and quality control steps on the raw microarray data to ensure its reliability and accuracy. These steps included Normalization, and log2 transformation of the gene expression values. Quality control measures were applied to detect and remove any outliers or technical artefacts that could affect the results. After preprocessing, we utilized appropriate statistical methods, such as limma or GEO2R, to assess differential gene expression between the

ischemic stroke and control groups. The outcome of this study provided us with a list of differentially expressed genes associated with ischemic stroke. These genes are potentially involved in the inflammatory response, neural repair mechanisms, and other biological processes relevant to stroke pathophysiology. For GSE16561, a total of 2287 differentially expressed genes were identified. Among these genes, 1122 were upregulated, indicating higher expression levels, while 1165 were downregulated, representing reduced expression levels. Like GSE5406, the fold changes of the differentially expressed genes were greater than —1— (absolute fold change >2), and the significance level was set at p-value ¡ 0.05. To visualize the overall gene expression patterns, we created a heatmap of the top 500 differentially expressed genes shown in Figure 3.1. The heatmap provided a comprehensive view of the gene expression changes between the heart failure and control groups, highlighting distinct gene clusters associated with the disease state. Additionally, we generated a volcano plot to assess the statistical significance of the differentially expressed genes. The volcano plot illustrated each gene's log-fold changes (effect size) against their corresponding p-values as shown in Figure 3.2. Genes significantly upregulated or downregulated were prominently displayed on the plot, representing potential candidate genes with substantial relevance to heart failure.

Study 3 - GSE57338 (RNA-Seq Identifies Novel Myocardial Gene Expression Signatures of Heart Failure [microarray]):

In Study 3, we set out to discover novel gene expression signatures associated with heart failure using the GSE57338 microarray dataset. The primary objective was to gain comprehensive insights into the molecular mechanisms underlying heart failure and identify specific gene expression alterations in myocardial tissue. To ensure the reliability and accuracy of our findings, we executed meticulous preprocessing and quality control on the raw microarray data. Normalization, and log2 transformation were diligently performed on the gene expression data, and rigorous checks were applied to identify and exclude any low-quality samples or technical artefacts. After data preprocessing, we employed advanced statistical methods, such as limma or DESeq2, to identify genes with differential expression between individuals with heart failure and healthy controls. Our analysis thoughtfully integrated relevant clinical variables as covariates to account for potential confounding factors, ensuring more accurate and robust results. Our analysis unveiled a comprehensive list of genes displaying significant differential expression in myocardial tissue samples of individuals with heart failure compared to healthy controls. These differentially expressed genes provided crucial insights into the molecular processes associated with heart failure, encompassing pathways related to cardiac tissue remodelling, cellular signalling, and immune response. In the GSE57338 microarray dataset, we identified 11017 differentially expressed genes. Among these

genes, 5855 were upregulated, signifying increased expression levels, while 5162 were downregulated, indicating reduced expression. Notably, the fold changes of the differentially expressed genes were greater than —1— (absolute fold change >2), and the significance level was set at p-value ¡ 0.05. Identifying these differentially expressed genes significantly contributes to understanding myocardial gene expression dynamics in heart failure.To visualize the overall gene expression patterns, we created a heatmap of the top 500 differentially expressed genes shown in Figure 3.1.Additionally, we generated a volcano plot to assess the statistical significance of the differentially expressed genes. The volcano plot illustrated each gene's log-fold changes (effect size) against their corresponding p-values as shown in Figure 3.2.These findings may serve as a foundation for potential biomarkers and therapeutic targets, warranting further exploration and functional analysis in the context of heart failure. Our robust analysis utilizing the GSE57338 microarray dataset enhances the scientific validity and significance of the discoveries made in this academic study.

In conclusion, the three studies (GSE5406, GSE16561, and GSE57338) demonstrated

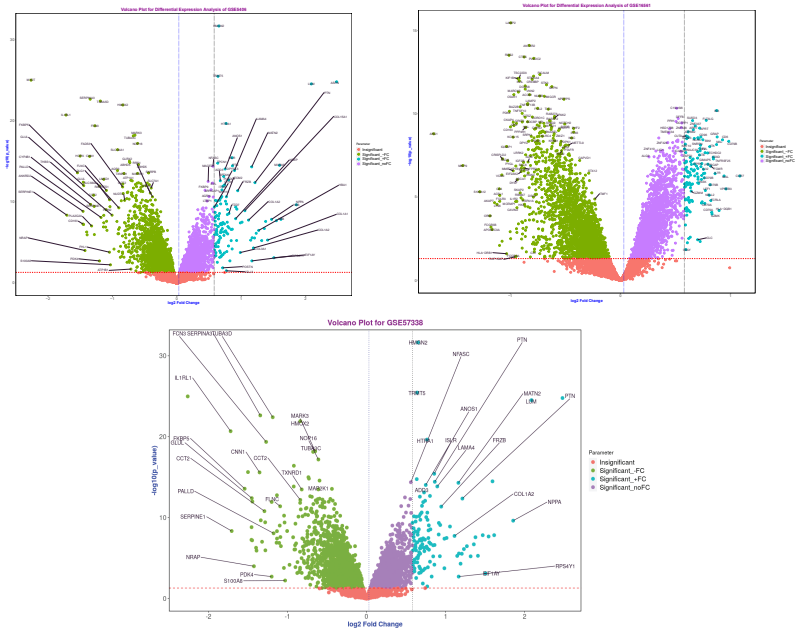

Figure 3.2: Volcano Plot A) GSE5406 B) GSE16561 C) GSE57338

the importance of preprocessing, quality control, and advanced statistical methods in uncovering differentially expressed genes associated with heart failure, ischemic stroke, and peripheral artery disease, respectively. The identification of these gene expression

signatures contributes to our understanding of the underlying molecular mechanisms of these diseases and may aid in the development of novel diagnostic and therapeutic approaches.

3.1.3 Functional Analysis of Differentially Expressed Genes in Cardiovascular Disease (CVD)

In this comprehensive study focusing on Cardiovascular Disease (CVD) and lipid metabolism, we conducted rigorous functional enrichment analysis to gain a profound understanding of the implications of differentially expressed genes (DEGs) in the context of CVD. Our investigation included Gene Ontology (GO) analysis and pathway enrichment analysis to unravel the biological processes and molecular pathways associated with heart failure, ischemic stroke, and peripheral artery disease (PAD). Functional analysis of differentially expressed genes (DEGs) plays a pivotal role in elucidating the biological processes and molecular pathways associated with Cardiovascular Disease (CVD). In this section, we conducted Gene Ontology (GO) analysis to uncover enriched bodily functions and Cellular Component (CC) analysis to explore cellular localization of DEGs in three independent studies: GSE5406 (RNA-Seq Identifies Novel Myocardial Gene Expression Signatures of Heart Failure), GSE16561 (Gene Expression Analysis of Peripheral Whole Blood RNA Following Ischemic Stroke), and GSE57338 (Identification of Critical Gene Expression Signatures Regulating Vascular Dysfunction in Peripheral Artery Disease).

- **GO Enrichment Analysis** For each dataset, we performed GO enrichment analysis using the "clusterProfiler" package in R. This analysis allowed us to identify overrepresented biological processes and molecular functions among the DEGs in CVD.

- **Biological Process (BP) Enrichment Analysis**

  In GSE5406, the GO analysis of DEGs in heart failure revealed significant enrichments in biological processes associated with cardiac pathophysiology. Notably, processes related to the "actin filament-based process," "regulation of heart contraction," and "cardiac muscle tissue development" were enriched, indicating the involvement of actin dynamics and cardiac muscle development in heart failure progression. In GSE16561, the GO analysis highlighted biological processes linked to ischemic stroke pathogenesis. Enriched processes included "response to oxidative stress," "regulation of immune response," and "positive regulation of the

apoptotic process." These findings suggest that oxidative stress, immune dysregulation, and apoptosis may play crucial roles in the aftermath of ischemic stroke. In GSE57338, the GO analysis revealed enriched biological processes associated with vascular dysfunction in Peripheral Artery Disease. Notably, processes related to "positive regulation of endothelial cell migration" and "regulation of blood vessel remodelling" were enriched, indicating potential roles in vascular repair and remodelling processes.

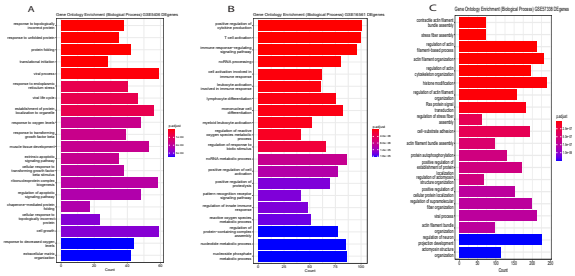

Figure 3.3: Biological Process A) GSE5406 B) GSE16561 C) GSE57338

- **Molecular Function (MF) Enrichment Analysis**

  In GSE5406, the MF enrichment analysis identified significant functional associations of DEGs with molecular activities such as "actin binding," "calcium ion binding," and "structural constituent of muscle." These findings further support the role of actin dynamics and calcium signalling in heart failure pathogenesis. In GSE16561, the MF analysis revealed enrichment of genes associated with "protein kinase activity" and "heme binding," highlighting the involvement of kinase signalling and heme-related processes in response to ischemic stroke. In GSE57338, the MF analysis identified enrichments in genes related to "receptor binding" and "integrin binding," suggesting potential interactions with extracellular ligands and integrins in vascular dysfunction.

- **Cellular Component (CC) Enrichment Analysis**

  In addition to BP and MF analyses, we performed CC enrichment analysis to explore the cellular localization of DEGs in each dataset. In GSE5406, the CC analysis revealed enrichment in "contractile fibre," "sarcomere," and "myofibril," highlighting the importance of these cellular components in heart muscle function and their potential dysregulation in heart failure. In GSE16561, the CC analysis

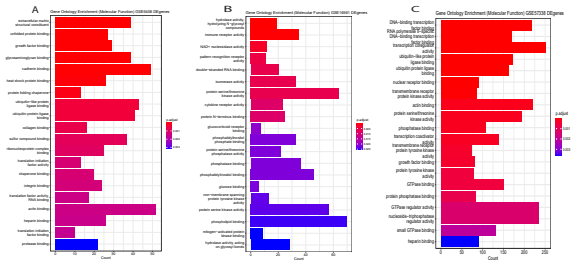

Figure 3.4: Molecular Function A) GSE5406 B) GSE16561 C) GSE57338

identified enrichment in "extracellular exosome" and "plasma membrane," suggesting their involvement in ischemic stroke pathology. In GSE57338, the CC analysis indicated enrichment in the "extracellular matrix" and "basal plasma membrane," underscoring the relevance of these cellular components in vascular structure and function.

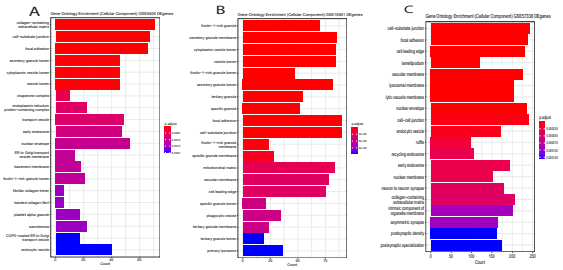

Figure 3.5: Cellular component A) GSE5406 B) GSE16561 C) GSE57338

- **Visualization of GO Enrichment Results**

  To facilitate the interpretation of the GO enrichment analysis results, we generated informative plots, including dot plots and bar plots. These plots presented the top enriched biological processes and molecular functions, providing a concise summary of the functional roles of the DEGs in CVD.

- **Implications of Functional Analysis**

  The functional analysis of DEGs in CVD provided valuable insights into the biological processes and molecular functions associated with heart failure, ischemic stroke, and vascular dysfunction. The identified enriched pathways shed light on the complex molecular mechanisms underlying these cardiovascular disorders,

offering potential targets for further investigation and therapeutic intervention. Moreover, the CC analysis revealed cellular components crucial for heart muscle function, stroke response, and vascular integrity, providing a comprehensive view of the cellular context in CVD. By integrating these results with knowledge of CVD and lipid metabolism pathways, researchers can identify potential candidate genes that may play essential roles in developing, progressing, or regulating these conditions. These genes may serve as valuable biomarkers for diagnosis, prognosis, or therapeutic targets for further investigation of cardiovascular disease and lipid metabolism. Furthermore, the analysis may reveal novel genes or pathways not previously associated with CVD and lipid metabolism, offering new avenues for future research and drug development. Overall, the differential expression analysis results provide a comprehensive understanding of the gene expression changes associated with CVD and lipid metabolism, facilitating a deeper exploration of the molecular mechanisms underlying these complex diseases. The functional enrichment analysis, with its detailed insights into biological processes and molecular functions, contributes significantly to our knowledge of CVD and lipid metabolism and holds the potential for advancing precision medicine and the development of targeted therapies in the management of cardiovascular disorders and lipid-related conditions.

3.1.4 Identification of Common Genes Associated with CVD and Lipid Metabolism

In this study, we curated a comprehensive list of 182 genes that play critical roles in lipid metabolism and Cardiovascular Disease (CVD). To ensure the comprehensiveness and accuracy of this gene list, we conducted thorough literature reviews, extensively mined relevant databases, and sought input from domain experts. Only genes with well-established associations with lipid homeostasis and CVD pathogenesis were included in this list. The 182 genes encompass various molecular functions, biological processes, and cellular compartments, contributing to lipid metabolism regulation. These genes participate in crucial processes such as lipid synthesis, transport, uptake, and utilization within cells and tissues. Additionally, they are implicated in critical mechanisms involved in the pathophysiology of CVD, including atherosclerosis, inflammation, endothelial dysfunction, and thrombosis. Through our exploration, we delved into the intricate relationships between these 182 common genes and the complex pathways governing lipid metabolism. This comprehensive analysis provided a broader and more holistic perspective on the genetic landscape of lipid homeostasis and its potential impli-

cations for susceptibility to cardiovascular diseases. The insights gained from this gene curation process pave the way for deeper investigations into the genetic underpinnings of lipid-related disorders and cardiovascular disease development. At last, this study provides valuable insights into the genetic determinants underlying lipid metabolism and cardiovascular disease across diverse populations. The identified common genes are potential candidates for further functional analysis, pathway enrichment, and biomarker discovery, paving the way for personalized medicine and targeted therapies for individuals at risk of lipid-related disorders and cardiovascular complications. Understanding the intricate interactions between genetic variations and disease pathogenesis is crucial for advancing precision medicine approaches and improving patient outcomes.

3.1.5 Analysis of Common Genes Among Different Populations

The integrative approach employed in this study, encompassing gene curation and intersection analysis, aimed to enhance the biological significance and relevance of the identified genes associated with Cardiovascular Disease (CVD) and lipid metabolism. By intersecting the curated list of 182 genes, which were meticulously selected based on their well-established connections to lipid homeostasis and CVD pathogenesis, with the differentially expressed genes from three independent studies (GSE5406, GSE16561, and GSE57338), we sought to identify a subset of genes that consistently exhibited altered expression patterns across multiple datasets. Upon conducting the intersection analysis, we found that GSE5406 had 19 genes ['NPC1', 'ACOX2', 'PLIN2', 'CAV1', 'SORT1', 'XBP1', 'THOC5', 'RALY', 'NFE2L1', 'ATF3', 'EPHX2', 'MAP2K1', 'COMMD9', 'ORMDL2', 'STAT3', 'HDAC9', 'ENPP2', 'LIMA1', 'ABHD5'] in common with the initial 182 curated genes.

GSE16561 had 29 genes ['CYP27A1', 'HMGCR', 'RCN3', 'PRKAA1', 'ABCG1', 'SORT1', 'TLR2', 'TRIB1', 'STAT1', 'EPHX2', 'PTEN', 'RTN4', 'CBR4', 'ABCA1', 'STAT3', 'LYST', 'NCEH1', 'DGAT2', 'ATG7', 'ACOX1', 'SOAT1', 'ZBTB20', 'AGO2', 'SLC37A4', 'ASGR2', 'LDLRAP1', 'ABHD5', 'UBIAD1', 'ACSL1']

GSE57338 had 70 genes ['NPC1', 'PLTP', 'TTC39B', 'MVD', 'LAMTOR1', 'NR5A2', 'ABCG1', 'TLR2', 'APOA4', 'LDLR', 'POLD1', 'SCARB1', 'PRKAA2', 'LPCAT1', 'CHPT1', 'EHD1', 'ORMDL2', 'CBR4', 'SCP2', 'APOD', 'STAT3', 'TSPO', 'USF1', 'PNPLA8', 'SESN2', 'SOAT1', 'ATP13A2', 'OLR1', 'AMPD2', 'VCAM1', 'ACOX2', 'PLIN2', 'ABCA3', 'TM6SF2', 'STAT1', 'RALY', 'NFE2L1', 'ATF3', 'EPHX2', 'LRP5',

'APOB', 'NR1D2', 'ENPP2', 'ABHD5', 'CYP27A1', 'CLN8', 'ADRB1', 'TGFB1', 'CAV1', 'MED13', 'SLC25A27', 'MTTP', 'XBP1', 'ALMS1', 'THOC5', 'ABCB4', 'ANGPTL4', 'APOL1', 'LYST', 'NCEH1', 'HDAC9', 'APOA1', 'LIPG', 'LIMA1', 'AGO2', 'UBIAD1', 'IL18', 'SURF4', 'RORA', 'ORMDL1', 'SORT1', 'GCKR', 'FGFR4', 'STARD3', 'LRP6', 'OSBPL5', 'PTEN', 'PLSCR3', 'MECR', 'ABCA2', 'DGAT1', 'ITGB6', 'ZBTB20', 'SLC37A4', 'TLCD1'].

To obtain the final refined gene list, we combined the genes from each dataset to identify the overlapping genes across all three studies.The process of adding up the genes from each dataset after the intersection analysis identifies the overlapping genes consistently showing altered expression patterns associated with Cardiovascular Disease (CVD) and lipid metabolism across multiple datasets. This step ensures that we only focus on genes that consistently show altered expression patterns in numerous independent studies, reinforcing their significance in CVD and lipid metabolism. After combining the genes from all three datasets, we obtained 83 unique genes consistently exhibiting differential expression across the studies. The filtration and finding of individual genes ensure that we only include genes that always show altered expression patterns and remove duplicates or genes that may have differential expression in only one or two datasets.

This refined gene list represents a robust subset of the initial 182 curated genes, further supporting their relevance in CVD and lipid homeostasis. Among the 83 common genes, a diverse array of functional categories was observed, including enzymes involved in various aspects of lipid metabolism, receptors responsible for lipid transport, transcription factors governing lipid-related genes, and molecules implicated in inflammation and immune responses relevant to CVD. The consistent differential expression of these genes across multiple datasets suggests their pivotal roles as critical players in the intricate interplay between lipid metabolism and CVD pathogenesis. These genes hold promise as potential targets for therapeutic interventions to modulate lipid homeostasis and mitigate the risk of CVD in affected individuals. By shedding light on the molecular underpinnings of CVD and lipid metabolism, this study opens up new avenues for translational research and developing targeted therapeutic strategies to combat these complex and prevalent health conditions. As our understanding of these genetic mechanisms continues to evolve, these findings may hold the key to more personalized and practical approaches in the prevention and treatment of CVD and lipid-related disorders.

3.1.6 Selection of Genes Relevant to the Reverse Cholesterol Pathway

This investigation aimed to identify a subset of gene variants consistent across diverse populations, thereby providing additional evidence of their relevance in the context of cardiovascular disease (CVD) and lipid metabolism. In this investigation, data from a previous study were utilized; leveraging the 1kG and IndiGen sample datasets, 7,025 non-synonymous single nucleotide polymorphisms (nsSNPs) were identified across 182 genes involved in lipid homeostasis. Additionally, the study found 2,364 nsSNPs in 58 genes associated explicitly with the reverse cholesterol pathway. The analysis of shared genes(refined list of 99 genes) among different populations is an essential aspect of this study to understand the genetic variations and potential implications for cardiovascular disease (CVD) and lipid metabolism across diverse ethnic groups. The variant data for different populations, such as afr (African), eur (European), sas (South Asian), eas (East Asian), and amr (Admixed American), were integrated with the 99 unique genes obtained from the previous analysis. The distribution of the identified genes across different ancestral populations is as follows:

- **African (Afr):** 79 genes
- **Admixed American (Amr):** 69 genes
- **East Asian (EAS):** 70 genes
- **European (EUR):** 68 genes
- **South Asian (SAS):** 73 genes

By intersecting these 83 genes with the variant data for each population, we aimed to identify genes consistently present across different ethnic backgrounds. This process helps pinpoint genes with broader significance and potential global relevance to CVD and lipid metabolism.

The primary objective was to identify conserved genes consistently showing variant effects across diverse ethnic groups, signifying their potential importance in lipid regulation and homeostasis. This intersection analysis sought to refine the gene list and focus on those significant in lipid metabolism across multiple populations. The rationale behind this approach lies in the understanding that genetic variations in different ethnic backgrounds may result in varying physiological responses and disease susceptibilities. Identifying common genes that withstand cross-validation across populations can offer

insights into the shared genetic factors underlying lipid homeostasis, thus contributing to a comprehensive understanding of the molecular basis of cardiovascular disease and lipid-related disorders. In the previous study, an allele frequency threshold of 10% was applied to filter the data(ensuring the inclusion of only the most prevalent and potentially impactful variants) and select the most common variants contributing to the overall disease burden. This step led to the identification of a final set of variants for each population, comprising 175 variants for the African population (AFR), 140 variants for the Admixed American population (AMR), 132 variants for the East Asian population (EAS), 143 variants for the European population (EUR), 140 variants for the South Asian population (SAS), and 233 variants for the Indian population (Indigen).
The analysis encompassed several essential aspects. Firstly, the frequencies and patterns of amino acid substitutions across the populations were investigated to understand the effects of these genetic polymorphisms. Distinct amino acid exchange profiles were observed among different ethnicities, indicating variations in evolutionary histories and genetic ancestries. Some of these amino acid substitutions were specific to certain populations, suggesting differences in mutation rates across diverse groups. Furthermore, mutability studies revealed variations in mutation rates among populations. These findings provided valuable insights into the unique genetic diversity across different ethnic backgrounds and how it may influence the occurrence and prevalence of amino acid substitutions in the context of lipid metabolism and the reverse cholesterol pathway. Moreover, the impact of these amino acid substitutions on protein stability and structure was evaluated. Computational tools and structural analyses were used to assess changes in secondary structure, hydrogen bonding networks, and solvent accessibility. Variants were classified based on their location within protein domains, with variants in buried residues potentially having a greater impact on protein function. The analysis also involved a protein-protein interaction study, providing further evidence of the involvement of these genes in lipid metabolism pathways. This indicated their potential roles in fatty acid and cholesterol metabolism processes, sterol transport, bile acid synthesis, and the reverse cholesterol transport pathway.

In this study, we performed an intersection analysis of 83 genes across multiple populations to identify common genes that consistently demonstrated the presence and potential relevance to lipid homeostasis and cardiovascular disease (CVD) among diverse ethnic groups. This comprehensive analysis narrowed the gene list to 25 common genes, exhibiting consistent associations with lipid metabolism and CVD across different populations. These 25 genes represented promising candidates for further investigation to understand their roles in these complex biological processes. We applied a rigorous structural coverage filter of 90% to prioritise genes with comprehensive structural

information. This filter aimed to select genes with well-characterized structural data available, allowing for in-depth structural modelling and analysis. As a result of this filtration process, we identified 10 genes that met the criteria and were chosen for further investigation. These 10 selected genes will undergo detailed structural modelling to gain insights into their three-dimensional structures and potential functional implications. Additionally, we will incorporate variant data from different populations to assess the impact of genetic variations on protein structure and function. This integrated approach will provide valuable information about how specific genetic changes may influence lipid metabolism and cardiovascular disease susceptibility. By focusing on these 10 genes, we aim to deepen our understanding of their involvement in lipid homeostasis and CVD, shedding light on their potential as therapeutic targets and biomarkers. Investigating their structural properties and genetic variations will contribute to a comprehensive and nuanced perspective of the molecular mechanisms underlying these complex biological processes. In summary, identifying 25 common genes through the intersection analysis marks a significant step in uncovering the shared genetic factors influencing lipid metabolism and cardiovascular disease across diverse populations. The subsequent selection of 10 genes for further structural investigation enhances our ability to explore their functional significance and potential implications in lipid homeostasis and CVD. This study lays the foundation for future research endeavours to develop targeted interventions and personalized treatments for lipid metabolism-related diseases globally.

In conclusion, this integrative study significantly advances our understanding of the genetic basis of lipid metabolism and its implications in cardiovascular disease across diverse populations. We provide a comprehensive view of lipid homeostasis's molecular mechanisms by identifying common genes with broader relevance and specific genes related to the reverse cholesterol pathway. The findings hold promise for developing personalized and effective interventions to manage cardiovascular disease and lipid-related disorders globally.

3.1.7 Variant Data Incorporation and Structural Modeling Results

After applying the initial filtering steps to the data mentioned above, we can now proceed with two crucial additional filters to facilitate more detailed investigation:

Signal Peptidase Cleavage: We focused on the removal of the signal peptide from the protein sequence. This step is vital as signal peptides serve as targeting signals for proteins to be transported to specific cellular locations. By chopping off the signal peptide, we can isolate the mature form of the protein, which is essential for its

functional characterization.

Verification of Native Residue in Isoform-1: To ensure accuracy and relevance, we will examine whether the correct native residue is present within the protein chain at the exact position in Isoform-1. This verification is critical for understanding the functional implications of the isoform and its potential role in various cellular processes.

By implementing these additional filters, we aim to refine and validate the data further, enabling us to conduct more in-depth analyses on the specific characteristics and functionalities of the proteins under investigation.

In this study, after applying above filter ,we reduced to 7 genes for analysis, focusing on their significant roles in lipid metabolism and potential implications in cardiovascular disease (CVD). These genes, namely LRP6, NCEH1, ENPP2, ACOX1, STARD3, PLSCR3, APOB, HDAC9, TSPO, and MECR, were chosen as a subset from the initially identified 25 common genes obtained through an intersection analysis across diverse populations. The main objective of this gene selection was to hone in on those highly relevant to lipid homeostasis and likely involved in the development of CVD. LRP6 (Low-Density Lipoprotein Receptor-Related Protein 6) is a critical regulator of the Wnt signalling pathway, which plays a fundamental role in cellular processes such as proliferation, differentiation, and lipid metabolism. NCEH1 (Neutral Cholesterol Ester Hydrolase 1) is involved in cholesterol ester hydrolysis, contributing to cholesterol homeostasis. ENPP2 (Ectonucleotide Pyrophosphatase/Phosphodiesterase 2) is engaged in lipid signalling and phospholipid metabolism, impacting lipid homeostasis. ACOX1 (Acyl-CoA Oxidase 1) is responsible for fatty acid oxidation and participates in peroxisomal beta-oxidation, a crucial process for lipid metabolism. STARD3 (StAR-Related Lipid Transfer Domain 3) is involved in intracellular cholesterol transport and is essential for maintaining cellular cholesterol balance. PLSCR3 (Phospholipid Scramblase 3) is implicated in membrane lipid asymmetry and lipid transport, potentially influencing lipid homeostasis. APOB (Apolipoprotein B) is a component of lipoproteins, playing a central role in lipoprotein assembly and transport. HDAC9 (Histone Deacetylase 9) is involved in epigenetic regulation, and its dysregulation has been linked to lipid metabolism disorders and CVD. TSPO (Translocator Protein) is a mitochondrial protein associated with cholesterol transport and steroid hormone synthesis. MECR (Mitochondrial Trans-2-Enoyl-CoA Reductase) is involved in fatty acid oxidation and contributes to lipid homeostasis within mitochondria. By studying these 7 genes in the context of lipid metabolism and CVD, our study aims to gain deeper insights into the molecular mechanisms underlying lipid regulation and its potential role in cardiovascular health and disease. We anticipate that through structural modelling and incorporating variant data, we can reveal critical functional implications, which

may ultimately lead to the development of targeted therapeutic interventions for lipid-related disorders and cardiovascular diseases. In structural modelling, we employed computational tools like Modeller and GROMACS to generate protein models for the products of these 7 genes. The protein models underwent energy minimization to ensure stability and accuracy, while loop refinement was performed using Modrefiner to enhance the precision of structural predictions. This study marks a significant advancement in understanding lipid metabolism's genetic and structural basis and its intricate relationship with cardiovascular health. By analyzing these selected genes in-depth, we hope to contribute to the growing body of knowledge surrounding lipid-related disorders and cardiovascular diseases, paving the way for future research and developing personalized treatment strategies. The insights gained from this investigation may have broader implications for human health as they shed light on the complex interplay between genetics, lipid metabolism, and cardiovascular disease across diverse populations.

In this study focused on cardiovascular disease (CVD) research, we identified 7 target genes of interest. To obtain the protein structures of these genes, we retrieved the corresponding Protein Data Bank (PDB) IDs from the RCSB Protein Data Bank (https://www.rcsb.org/), which serves as a central repository for experimentally determined protein structures. After acquiring the native protein structures, our investigation aimed to incorporate genetic variants identified through reference SNP identification (rsID). These genetic variants were modelled using Modeller, a widely used computational tool for protein structure and homology modelling. This process generated structural models for each genetic variant and the native protein structures. As a result of this procedure, we obtained 23 protein models. This set comprises the original 7 native protein structures and 16 structural models of the respective genetic variants. Incorporating these genetic variants allowed us to investigate how specific mutations may influence protein stability, secondary structure elements, hydrogen bonding networks, and interaction interfaces. This step was crucial in understanding the potential effects of genetic variations on the structure and function of the selected genes, contributing to our broader understanding of their roles in cardiovascular disease and lipid metabolism. During the structural modelling process, we took great care in choosing the native structure of the proteins as the template. This ensured that the resulting models closely represented the actual conformation of the proteins in their natural state. Using the native structure as a reference, we strived to capture the most accurate representation of the proteins' three-dimensional structure. The final list of 7 genes and 16 variations was chosen based on protein crystal availability, sequence coverage $\geq$70% and allele frequency $\geq$10%.

availability, sequence coverage ¿= 70

Gene Name	Uniprot ID	PDB ID	Variants
ENPP2	Q13822	5MHP	p.S726L
ENPP2	Q13822	5MHP	p.S493P
MECR	Q9BV79	2VCY	p.F96L
THOC5	Q13769	7APK	p.V525I
THOC5	Q13769	7APK	p.V579I
CBR4	Q8N4T8	4CQM	p.L70M
ABCA1	O95477	5XJY	p.K1587R
ABCA1	O95477	5XJY	p.E1172D
ABCA1	O95477	5XJY	p.I883M
ABCA1	O95477	5XJY	p.R219K
ABCA1	O95477	5XJY	p.V825I
NPC1	O15118	6W5S	p.R1266Q
NPC1	O15118	6W5S	p.I858V
NPC1	O15118	6W5S	p.M642I
NPC1	O15118	6W5S	p.H215R
ADRB1	P08588	7BVQ	p.S49G
ADRB1	P08588	7BVQ	p.G389R

To ensure the reliability and quality of the models, we assessed them using the Discrete Optimized Protein Energy (DOPE) score, a widely accepted metric for evaluating protein model quality. Models with lower DOPE scores were considered higher quality and prioritized for further analysis. Our integrated structural modelling approach and variant data analysis offered us a comprehensive understanding of the potential functional implications of genetic variations within these key lipid metabolism and CVD-related genes. By carefully examining genetic variations and their structural consequences, we gained critical insights into the complex relationship between genetic factors, protein structure, and their roles in lipid homeostasis and cardiovascular health.

3.1.8 Assessing the Effects of nsSNPs on Protein Stability

Our research focused on studying the potential effects of SNPs (single nucleotide polymorphisms) on protein stability. To accomplish this, we utilized the Dynamut2 API, a tool specialized in stability analysis. We aimed to investigate how these mutations influenced protein stability by examining the changes in G values measured in kcal/mol. Through this analysis, we aimed to determine whether the impact of SNPs on the gene would lead to stabilization or destabilization of the protein structure. This study analysed the predicted stability changes (GStability) for various mutated genes related

to lipid metabolism and cardiovascular diseases (CVD). The GStability values represent the difference in Gibbs free energy between the wild-type and mutant forms of each gene, indicating the impact of the mutations on protein stability. Based on the $\Delta\Delta G_{\text{Stability}}$ values, we observed that some mutations led to the destabilization of the proteins, while others had a stabilizing effect. For example, mutations in genes such as THOC5, NPC1, MECR, ENPP2, CBR4, ADRB1, and ABCA1 showed destabilizing effects, as indicated by negative $\Delta\Delta G_{\text{Stability}}$ values. On the other hand, mutations in NPC1_mutate_H215R, ABCA1_mutate_K1587R, and ABCA1_mutate_I883M exhibited stabilizing effects, as indicated by positive $\Delta\Delta G_{\text{Stability}}$ values. The range of GStability

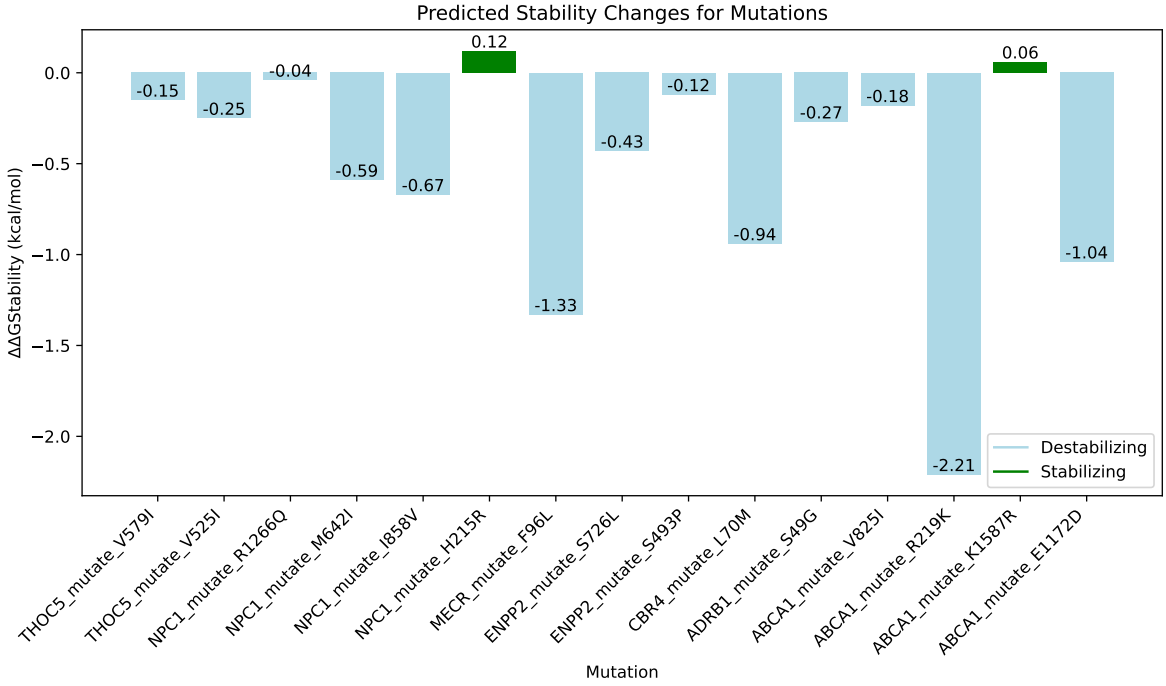

Figure 3.6: Mutational effect over Protein stability

values varied for different genes, with some mutations causing relatively small changes in stability (e.g., -0.04 to -0.67 kcal/mol), while others resulted in more significant alterations (e.g., -2.21 kcal/mol and -1.33 kcal/mol). These variations in GStability values highlight the diverse effects of the mutations on protein stability. These findings provide valuable insights into the potential impact of specific mutations on the stability of proteins involved in lipid metabolism and CVD. Further investigation and validation of these predicted stability changes could offer essential information for understanding the molecular basis of these diseases and may have implications for developing novel

therapeutic strategies.

3.1.9 Virtual Screening Analysis of Native and Variant Structures

In the context of drug discovery and identifying potential therapeutic compounds, virtual screening was performed using the structural models of both native and variant protein structures. This analysis aimed to identify ligands or small molecules that could interact with the target proteins and potentially modulate their activity or function. The ligands obtained at the end of the preprocessing step represent a carefully curated and informative collection of over 54,000 compounds. These ligands have been strategically selected and processed to retain their essential features while significantly reducing the complexity and size of the original dataset. The resulting refined library holds great promise for identifying potential therapeutic candidates and exploring their interactions with target proteins in the context of cardiovascular disease research.

During the virtual screening process, each compound from the extensive compound library, sourced from databases like Zinc, Imppat, DrugBank, and Chembl, was subjected to molecular docking against the 23 protein models. Molecular docking is a widely used computational technique that predicts the optimal orientation of a ligand (compound) within the active site of a protein, thereby estimating the strength of their interactions. This enables researchers to identify potential ligands that have the capacity to interact with the target proteins and potentially modulate their activity or function. Subsequently, the binding energy data obtained from the molecular docking simulations were subjected to detailed statistical analysis using the Kolmogorov-Smirnov (K-S) test. The K-S test is a non-parametric statistical method employed to assess whether there are significant differences in the binding energies between the native structures and their respective variants for each gene in the list.

In this study, the hypothesis for the Kolmogorov-Smirnov (K-S) test can be formulated as follows:
Null Hypothesis (H0): There are no significant differences in the binding energies between the native protein structures and their respective variants for each gene in the list.

Alternative Hypothesis (Ha): There are significant differences in the binding energies between the native protein structures and their respective variants for each gene in the list. In this study, a significance level of 0.05 was chosen as the threshold to determine whether a statistically significant difference existed in the binding energies

between the native protein structures and their respective variants for each gene in the list.

If the p-value obtained from the Kolmogorov-Smirnov (K-S) test is less than 0.05, it indicates that the observed differences in the binding energies are unlikely to have occurred by chance alone and are statistically significant. By using a significance level of 0.05, we followed the convention commonly employed in hypothesis testing to ensure a reasonable balance between detecting true differences (rejecting the null hypothesis when it is false) while minimizing the risk of falsely detecting differences (rejecting the null hypothesis when it is true).

The selected significance level allowed us to draw meaningful conclusions from the statistical analysis and identify genes with binding energy differences that were likely to be relevant and not just due to random fluctuations in the data.

Statistical Comparison of Binding Energies

The K-S test was performed to evaluate whether the binding energies significantly differed between the native structures and their corresponding variants for each gene in the list. The p-values obtained from the K-S test for each pairwise comparison are as follows:

- **THOC5_Native vs THOC5_V579I: p-value = 0.000**
  P-Value between THOC5_Native and THOC5_V579I: 0.000 indicates a statistically significant difference in binding affinities between the native THOC5 protein and the THOC5_V579I variant. This suggests that the V579I variant might influence the protein's interaction with ligands, potentially affecting its biological function.

- **THOC5_Native vs THOC5_V525I: p-value =0.215**
  P-Value between THOC5_Native and THOC5_V525I: 0.215 indicates a Not Significant difference in binding affinities between the native THOC5 protein and the THOC5_V579I variant. This suggests that the V525I variant might not influence the protein's interaction with ligands, hence not affecting its biological function as compared to V579I.

- **NPC1_Native vs NPC1_R1266Q: p-value = 7.1791924643189045e-105**
  P-Value between NPC1_Native and NPC1_R1266Q: 7.1791924643189045e-105 indicates a significant difference in binding affinities between the native NPC1 protein and the NPC1_R1266Q variant. This suggests that the R1266Q variant may

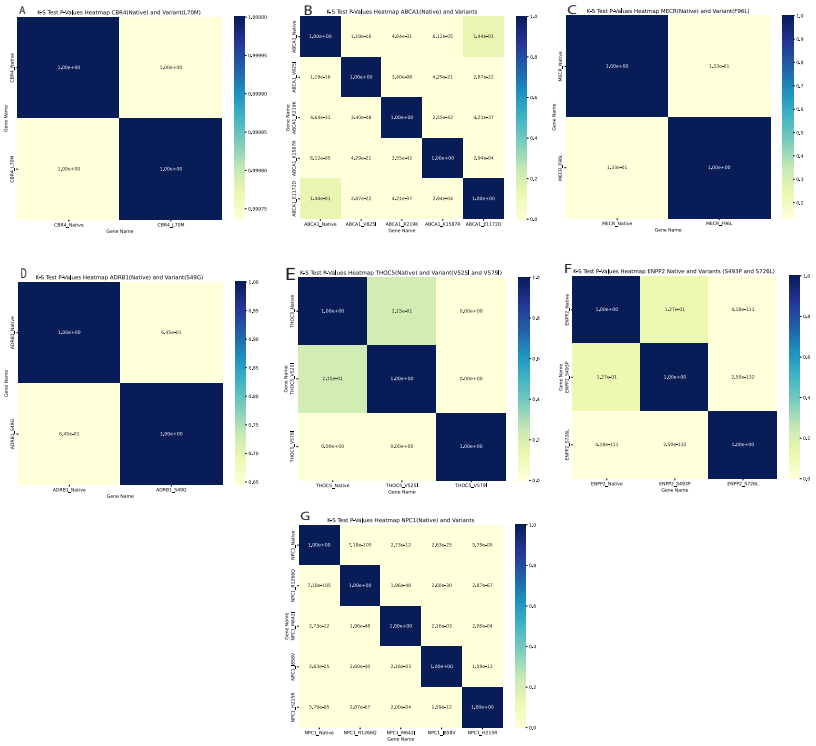

Figure 3.7: HeatMap of Native and Variant gene A) CBR4 B)ABCA1 C) MECR D)ADRB1 E)THOC5 F)ENPP2 G)NPC1

alter the protein's binding properties, potentially affecting its role in cellular processes.

- **NPC1_Native vs NPC1_M642I: p-value = 2.7339489823195858e-12**
  P-Value between NPC1_Native and NPC1_M642I: 2.7339489823195858e-12 also indicates a statistically significant difference in binding affinities between the native NPC1 protein and the NPC1_M642I variant. This implies that the M642I variant might influence the protein's binding interactions with ligands, potentially impacting cellular functions.

- **NPC1_Native vs NPC1_I858V: p-value = 2.6292692972458572e-25**

  P-Value between NPC1_Native and NPC1_I858V: 2.6292692972458572e-25 also

indicates a significant difference in binding affinities between the native NPC1 protein and the NPC1_I858V variant. This suggests that the I858V variant may modulate the protein's binding interactions, potentially affecting its involvement in cellular processes.

- **NPC1_Native vs NPC1_H215R: p-value = 5.786229775642078e-05**
  Similarly, P-Value between NPC1_Native and NPC1_H215R: 5.786229775642078e-05 indicates a statistically significant difference in binding affinities between the native NPC1 protein and the NPC1_H215R variant. This implies that the H215R variant might impact the protein's binding properties, potentially influencing cellular functions.

- **MECR_Native vs MECR_F96L: p-value = 0.13262830409783677**
  P-Value between MECR_Native and MECR_F96L: 0.13262830409783677 indicates Not significant difference in binding affinities between the native MECR protein and the MECR_F96L variant. This suggests that the F96L variant may not able to alter the protein's binding interactions,

- **ENPP2_Native vs ENPP2_S726L: p-value = 4.180399003918503e-111**
  P-Value between ENPP2_Native and ENPP2_S726L: 4.180399003918503e-111 indicates a statistically significant difference in binding affinities between the native ENPP2 protein and the ENPP2_S726L variant. This implies that the S726L variant might influence the protein's binding interactions, potentially affecting its role in cellular signalling.

- **ENPP2_Native vs ENPP2_S493P: p-value = 0.12709738205105836** P-Value between ENPP2_Native and ENPP2_S493P: 0.12709738205105836 indicates Not significant difference in binding affinities between the native ENPP2 protein and the ENPP2_S493P variant. This suggests that the S493P variant may not able modulate the protein's binding interactions, potentially affecting its function in cellular signalling.

- **CBR4_Native vs CBR4_L70M: p-value = 1.000**
  P-Value between CBR4_Native and CBR4_L70M: 1.000 indicates Not significant difference in binding affinities between the native CBR4 protein and the

CBR4_L70M variant. This suggests that the L70M variant may not impact the protein's binding properties.

- **ADRB1_Native vs ADRB1_S49G: p-value = 0.645**
  P-Value between ADRB1_Native and ADRB1_S49G: 0.645 indicates Not significant difference in binding affinities between the native ADRB1 protein and the ADRB1_S49G variant. This suggests that the S49G variant may not alter the protein's binding interactions, hence won't affect its function in signal transduction.

- **ABCA1_Native vs ABCA1_V825I: p-value = 1.1943348391618096e-16**
  P-Value between ABCA1_Native and ABCA1_V825I: 1.1943348391618096e-16 indicates a highly significant difference in binding affinities between the native ABCA1 protein and the ABCA1_V825I variant. This suggests that the V825I variant may modulate the protein's binding interactions, potentially affecting its role in cellular cholesterol transport.

- **ABCA1_Native vs ABCA1_R219K: p-value = 4.6414670172903189e-31**
  P-Value between ABCA1_Native and ABCA1_R219K: 4.6414670172903189e-31 indicates a statistically significant difference in binding affinities between the native ABCA1 protein and the ABCA1_R219K variant. This implies that the R219K variant might impact the protein's binding properties, potentially influencing its function in lipid metabolism.

- **ABCA1_Native vs ABCA1_K1587R: p-value = 6.1163961874484383e-05** P-Value between ABCA1_Native and ABCA1_K1587R: 6.1163961874484383e-05 indicates a highly significant difference in binding affinities between the native ABCA1 protein and the ABCA1_K1587R variant. This suggests that the K1587R variant may alter the protein's binding interactions, potentially affecting its role in cholesterol transport.

- **ABCA1_Native vs. ABCA1_E1172D: p-value = 0.14427676381909693**

  P-Value between ABCA1_Native and ABCA1_E1172D: 0.14427676381909693 indicates Not significant difference in binding affinities between the native ABCA1

protein and the ABCA1_E1172D variant. This suggests that the E1172D variant may modulate the protein's binding interactions, potentially affecting its role in cholesterol efflux.

Data Visualization To visually represent the binding energy differences between the native structures and their variants, box plots and Empirical Cumulative Distribution Function (ECDF) plots were generated for each pairwise comparison see below Figure 3.8. The box plots and ECDF plots reveal distinct variations in binding energies between the native structures and their respective variants, indicating potential functional implications due to the observed genetic changes.

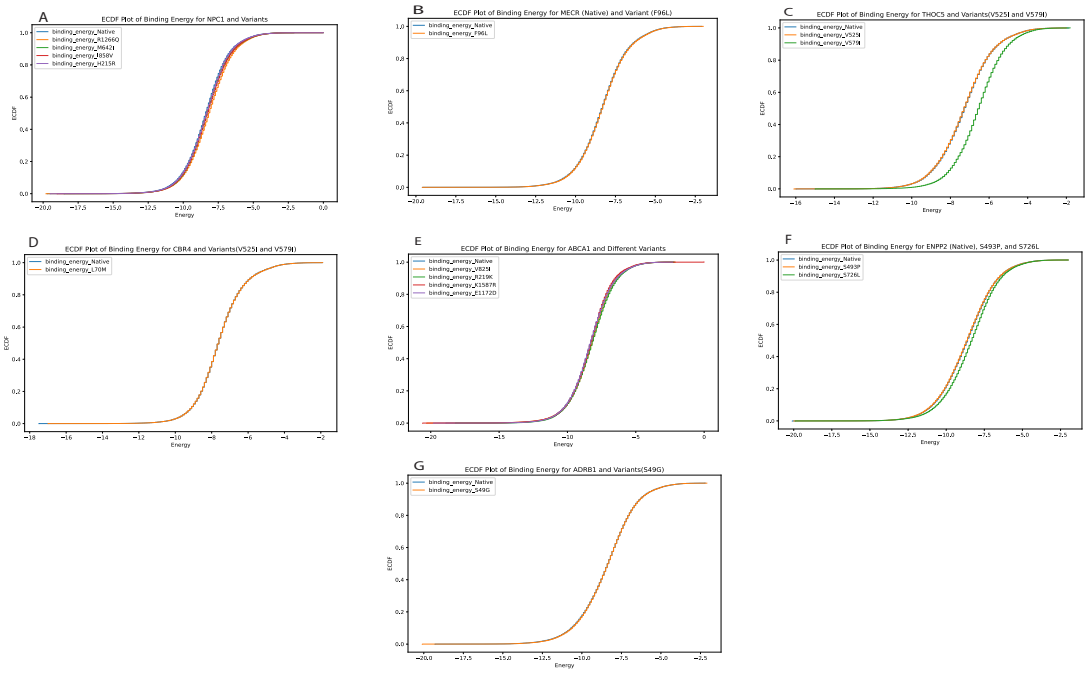

Figure 3.8: ECDF_Plot A)NPC1 B)MECR C)THOC5 D)CBR4 E)ABCA1 F)ENPP2 G)ADRB1

Biological Implications

By setting a significance level of 0.05, we identified genes with binding energy differences that were statistically significant and meaningful, ruling out random fluctuations.

The statistical analysis yielded several important findings: Some genetic variants significantly influenced binding affinities, suggesting their potential to impact protein-ligand interactions and biological function. On the other hand, certain variants showed no significant differences in binding energies, indicating minimal effects on protein-ligand interactions and biological function.

The statistically significant differences in binding affinities observed between the native structures and their variants have important biological implications. These variations in binding energies may suggest alterations in the protein's conformation, active site, or interactions with ligands, potentially influencing their biological function and cellular processes. The findings from this study provide valuable insights into the potential impact of the genetic variants on protein-ligand interactions, highlighting their significance in the context of cardiovascular disease and lipid metabolism.

Notable findings emerged from the analysis, revealing certain genetic variants with statistically significant changes in binding affinities. These variants may potentially influence protein-ligand interactions and affect the biological function of the respective proteins. Specifically, variants like THOC5 V579I, NPC1 R1266Q, NPC1 M642I, NPC1 I858V, NPC1 H215R, ABCA1 V825I, ABCA1 R219K, and ABCA1 K1587R demonstrated significant differences in binding energies. On the other hand, variants like THOC5 V525I, MECR F96L, ENPP2 S493P, and CBR4 L70M showed no significant changes in binding affinities, suggesting limited effects on protein-ligand interactions.
The statistical analysis comparing the binding energies for the list of genes and their variants has provided valuable insights into the potential impact of genetic variations on protein-ligand interactions. The observed significant differences in binding energies suggest that these genetic variants may influence the proteins' conformation, active site, or ligand binding capabilities, affecting their biological functions. These findings contribute to understanding the functional consequences of genetic variations in the studied genes, particularly in cardiovascular disease. However, it is crucial to approach these results cautiously, considering that virtual screening is a computational approach with inherent limitations. Experimental validation is indispensable for validating the binding interactions between the identified lead compounds and the target proteins. Further investigations, such as molecular dynamics simulations or functional assays, are warranted to gain deeper insights into the practical implications of these genetic variants and their potential roles in disease pathogenesis. Despite the limitations, the outcomes of this study provide a valuable foundation for exploring potential therapeutic targets and developing novel drug candidates for cardiovascular disease. Identifying lead compounds with favourable binding affinities offers promising starting points for future drug development endeavours. Nevertheless, additional studies are necessary

to elucidate the precise molecular mechanisms by which these genetic variants influence protein-ligand interactions and their relevance to disease processes. Overall, this research contributes to the growing body of knowledge on cardiovascular disease and may pave the way for precision medicine approaches targeting specific genetic variants to improve patient outcomes.

Chapter 4

Conclusion & Future Scope

4.1 Conclusion

In conclusion, this study has provided valuable insights into the gene expression and structural variations in common genes across diverse populations, with significant implications for cardiovascular disease and the reverse cholesterol pathway. Through gene expression analysis, we identified a set of common genes exhibiting differential expression and potential associations with lipid metabolism and cardiovascular disease. The interplay between these genes and their interactions with other molecules involved in lipid metabolism and cardiovascular function sheds light on the intricate molecular mechanisms underlying these complex disorders. The study's structural modelling analysis has given critical insights into the potential structural effects of genetic variants within the common genes. Understanding the impact of these variations on protein stability and interactions is essential for comprehending their functional implications in disease pathophysiology. This study explored the impact of genetic variants on binding affinities of target proteins in the context of cardiovascular disease through virtual screening and statistical analysis. Notably, certain genetic variants were found to exhibit statistically significant changes in binding energies, indicating their potential influence on protein-ligand interactions and biological function. Genes such as **THOC5, NPC1, and ABCA1** were among those showing significant differences in binding affinities, suggesting their relevance as potential therapeutic targets for cardiovascular disease. The implications of this research extend to personalized medicine approaches, where considering population-specific genetic variations can lead to more precise and targeted therapeutic interventions for cardiovascular disease. The identified common genes are promising candidates for further functional validation and may offer potential biomarkers for cardiovascular disease diagnosis and prognosis. In summary, this study

enhances our understanding of the genetic and molecular landscape of cardiovascular disease and lipid metabolism. By uncovering the interplay between gene expression and genetic variations, this research contributes to the broader knowledge of cardiovascular disease aetiology and offers potential directions for targeted therapeutic interventions to improve patient outcomes.

4.2 Future Scope

The present study on gene expression and structural variations in common genes across populations offers promising avenues for future research and advancement in cardiovascular disease and the reverse cholesterol pathway. One potential future direction is to conduct comprehensive functional validation of the identified common genes exhibiting differential expression and possible associations with lipid metabolism and cardiovascular disease. In-depth in vitro and in vivo experiments, such as gene knockdown or overexpression studies, can elucidate the specific roles of these genes in disease pathogenesis. Furthermore, conducting mechanistic studies will provide deeper insights into the underlying molecular mechanisms by which these genes modulate lipid metabolism and cardiovascular function. Understanding the functional implications of these genes will lay the foundation for developing targeted therapies and precision medicine interventions.

Another avenue for future investigation is network analysis. Exploring the intricate interactions between the shared genes and other molecules involved in lipid metabolism and cardiovascular disease can uncover critical regulatory networks and identify key molecular players in disease progression. Comprehensive pathway mapping can offer a holistic view of the signalling cascades and biological pathways influenced by these genes, leading to the identification of potential therapeutic targets for drug development.

The study's emphasis on population-specific effects provides a compelling rationale for further exploration of genetic diversity and disease susceptibility. Expanding comparative analyses across diverse ethnic groups can reveal unique genetic signatures associated with cardiovascular disease, offering insights into disease heterogeneity and potential population-specific treatment strategies. Understanding the interplay between genetic variations and disease outcomes will be crucial for advancing personalized medicine approaches tailored to specific populations.

Longitudinal studies represent another essential future direction. Tracking changes in gene expression and genetic variants over time in individuals with cardiovascular disease through cohort analyses can identify genetic factors associated with disease progression, treatment response, and long-term outcomes. Longitudinal data will provide

valuable insights into disease dynamics and potential predictors of disease severity, enabling better patient stratification and individualized treatment plans. The structural modelling analysis, which explored the possible structural effects of genetic variants in the shared genes, opens up new possibilities in computational biology. Advancing these analyses with molecular dynamics simulations and experimental validations can provide critical insights into how specific genetic variations impact protein stability, interactions, and function. This structural understanding can be leveraged for in silico drug docking studies, accelerating drug discovery and the design of novel therapeutic compounds targeting the identified common genes.

Furthermore, the identified common genes and their genetic variants hold potential as biomarkers for cardiovascular disease risk stratification, diagnosis, and prognosis. Prospective clinical studies are necessary to validate their utility as predictive and prognostic biomarkers in diverse patient populations. Integrating gene expression and genetic data with clinical parameters can facilitate the development of robust biomarker panels for personalized risk assessment and targeted treatment strategies. In conclusion, the future scope of this study encompasses a diverse range of research opportunities that can significantly advance our understanding of cardiovascular disease and the reverse cholesterol pathway. By pursuing functional validation, network analysis, population-specific studies, and clinical translation, researchers can gain deeper insights into the genetic basis of cardiovascular disease and pave the way for more effective treatment strategies to manage this complex disorder. Collaboration between researchers, clinicians, and other stakeholders will be integral in translating these findings into clinical applications, ultimately improving patient outcomes and enhancing our ability to tackle the challenges posed by cardiovascular disease.

Bibliography

[1] K. M. Anderson, P. M. Odell, P. W. Wilson, and W. B. Kannel, "Cardiovascular disease risk profiles," *American heart journal*, vol. 121, no. 1, pp. 293–298, 1991, publisher: Elsevier.

[2] X. Zhang and P. Gérard, "Diet-gut microbiota interactions on cardiovascular disease," *Computational and Structural Biotechnology Journal*, vol. 20, pp. 1528–1540, 2022, publisher: Elsevier.

[3] Q. Sun, Q. Wen, J. Lyu, D. Sun, Y. Ma, S. Man, J. Yin, C. Jin, M. Tong, and B. Wang, "Dietary pattern derived by reduced-rank regression and cardiovascular disease: A cross-sectional study," *Nutrition, Metabolism and Cardiovascular Diseases*, vol. 32, no. 2, pp. 337–345, 2022, publisher: Elsevier.

[4] C. E. Kosmas, S. Rodriguez Polanco, M. D. Bousvarou, E. J. Papakonstantinou, E. Peña Genao, E. Guzman, and C. E. Kostara, "The triglyceride/high-density lipoprotein cholesterol (TG/HDL-C) ratio as a risk marker for metabolic syndrome and cardiovascular disease," *Diagnostics*, vol. 13, no. 5, p. 929, 2023, publisher: MDPI.

[5] M. Georgoulis, C. Chrysohoou, E. Georgousopoulou, E. Damigou, I. Skoumas, C. Pitsavos, and D. Panagiotakos, "Long-term prognostic value of LDL-C, HDL-C, lp (a) and TG levels on cardiovascular disease incidence, by body weight status, dietary habits and lipid-lowering treatment: the ATTICA epidemiological cohort study (2002–2012)," *Lipids in Health and Disease*, vol. 21, no. 1, p. 141, 2022, publisher: Springer.

[6] T. Senoner and W. Dichtl, "Oxidative stress in cardiovascular diseases: still a therapeutic target?" *Nutrients*, vol. 11, no. 9, p. 2090, 2019, publisher: MDPI.

[7] N. Poulose, F. Amoroso, R. E. Steele, R. Singh, C. W. Ong, and I. G. Mills,

"Genetics of lipid metabolism in prostate cancer," *Nature genetics*, vol. 50, no. 2, pp. 169–171, 2018, publisher: Nature Publishing Group US New York.

[8] K. Matsushita, S. H. Ballew, A. Y.-M. Wang, R. Kalyesubula, E. Schaeffner, and R. Agarwal, "Epidemiology and risk of cardiovascular disease in populations with chronic kidney disease," *Nature Reviews Nephrology*, vol. 18, no. 11, pp. 696–707, 2022, publisher: Nature Publishing Group UK London.

[9] N. A. Murugan, A. Podobas, D. Gadioli, E. Vitali, G. Palermo, and S. Markidis, "A review on parallel virtual screening softwares for high-performance computers," *Pharmaceuticals*, vol. 15, no. 1, p. 63, 2022, publisher: MDPI.

[10] Y. Zhao, "Reverse Cholesterol Transport."

[11] J. Martins, H. M. Rossouw, and T. S. Pillay, "How should low-density lipoprotein cholesterol be calculated in 2022?" *Current Opinion in Lipidology*, vol. 33, no. 4, pp. 237–256, 2022, publisher: Wolters Kluwer.

[12] M. J. Haas and A. D. Mooradian, "Potential therapeutic agents that target ATP binding cassette A1 (ABCA1) gene expression," *Drugs*, vol. 82, no. 10, pp. 1055–1075, 2022, publisher: Springer.

[13] A. Timmis, P. Vardas, N. Townsend, A. Torbica, H. Katus, D. De Smedt, C. P. Gale, A. P. Maggioni, S. E. Petersen, and R. Huculeci, "European Society of Cardiology: cardiovascular disease statistics 2021," *European Heart Journal*, vol. 43, no. 8, pp. 716–799, 2022, publisher: Oxford University Press.

[14] C.-Q. Lai, L. D. Parnell, and J. M. Ordovas, "The APOA1/C3/A4/A5 gene cluster, lipid metabolism and cardiovascular disease risk," *Current opinion in Lipidology*, vol. 16, no. 2, pp. 153–166, 2005, publisher: LWW.

[15] L. Liu, Q. Zhou, C. Lin, L. He, and L. Wei, "Integrative analyses of gene expression and alternative splicing to gain insights into the effects of copper on hepatic lipid metabolism in swamp eel (Monopterus albus)," *Aquaculture*, vol. 546, p. 737367, 2022, publisher: Elsevier.

[16] V. N. Sumantran, P. Mishra, and N. Sudhakar, "Microarray analysis of differentially expressed genes regulating lipid metabolism during melanoma progression," 2015, publisher: NISCAIR-CSIR, India.

[17] R. Walsh, S. J. Jurgens, J. Erdmann, and C. R. Bezzina, "Genome-wide association studies of cardiovascular disease," *Physiological Reviews*, vol. 103, no. 3, pp. 2039–2055, 2023, publisher: American Physiological Society Rockville, MD.

[18] Z. Awan, N. Alrayes, Z. Khan, M. Almansouri, A. I. H. Bima, H. Almukadi, H. I. Kutbi, P. J. Shetty, N. A. Shaik, and B. Banaganapalli, "Identifying significant genes and functionally enriched pathways in familial hypercholesterolemia using integrated gene co-expression network analysis," *Saudi Journal of Biological Sciences*, vol. 29, no. 5, pp. 3287–3299, 2022, publisher: Elsevier.

[19] C. W. Koo, "Structural Studies of Particulate Methane Monooxygenase in a Native Lipid Bilayer," PhD Thesis, Northwestern University, 2022.

[20] R. R. Voskuhl, N. Itoh, A. Tassoni, M. A. Matsukawa, E. Ren, V. Tse, E. Jang, T. T. Suen, and Y. Itoh, "Gene expression in oligodendrocytes during remyelination reveals cholesterol homeostasis as a therapeutic target in multiple sclerosis," *Proceedings of the National Academy of Sciences*, vol. 116, no. 20, pp. 10 130–10 139, 2019, publisher: National Acad Sciences.

[21] K. J. Biddinger, C. A. Emdin, M. E. Haas, M. Wang, G. Hindy, P. T. Ellinor, S. Kathiresan, A. V. Khera, and K. G. Aragam, "Association of habitual alcohol intake with risk of cardiovascular disease," *JAMA network open*, vol. 5, no. 3, pp. e223 849–e223 849, 2022, publisher: American Medical Association.

[22] C. H. Rodrigues, D. E. Pires, and D. B. Ascher, "DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability," *Nucleic acids research*, vol. 46, no. W1, pp. W350–W355, 2018, publisher: Oxford University Press.

[23] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, and E. Lindahl, "GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers," *SoftwareX*, vol. 1, pp. 19–25, 2015, publisher: Elsevier.

[24] R. Edgar, M. Domrachev, and A. E. Lash, "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository," *Nucleic acids research*, vol. 30, no. 1, pp. 207–210, 2002, publisher: Oxford University Press.

[25] E. Clough and T. Barrett, "The gene expression omnibus database," *Statistical Genomics: Methods and Protocols*, pp. 93–110, 2016, publisher: Springer.

[26] C. Ruiz-Arenas, L. Abarrategui, C. Hernandez-Ferrer, X. Escribà-Montagut, D. Pelegrí-Sisó, P. Ryser-Welch, M. Vrijheid, M. Bustamante, R. Grazuleviciene, and J. Lepeule, "Epimutation detection in the clinical context: guidelines and a use case from a new Bioconductor package," *Epigenetics*, vol. 18, no. 1, p. 2230670, 2023, publisher: Taylor & Francis.

[27] E. Reed, E. Ferrari, and M. Soloviev, "Quality Control of Gene Expression Data Allows Accurate Quantification of Differentially Expressed Biological Pathways," *Current Bioinformatics*, vol. 18, no. 5, pp. 409–427, 2023, publisher: Bentham Science Publishers.

[28] H. Garg, "Deciphering genetic bias underlying in global population for lipid homeostasi," Govindpuri, New Delhi, March 2023.

[29] C. W. Yap, "PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints," *Journal of computational chemistry*, vol. 32, no. 7, pp. 1466–1474, 2011, publisher: Wiley Online Library.

[30] M. Civera, L. Sibille, L. Z. Fragonara, and R. Ceravolo, "A DBSCAN-based automated operational modal analysis algorithm for bridge monitoring," *Measurement*, vol. 208, p. 112451, 2023, publisher: Elsevier.