Deciphering Genetic Bias Underlying in Global Population for Lipid Homeostasis

A Project Report

submitted by

HIMANSHI GARG

in partial fulfilment of the requirements
for the award of the degree of

MASTER OF TECHNOLOGY

COMPUTATIONAL BIOLOGY
INDRAPRASTHA INSTITUTE OF INFORMATION TECHNOLOGY DELHI

NEW DELHI- 110020

15th March 2023

THESIS CERTIFICATE

This is to certify that the thesis titled **Deciphering Genetic Bias Underlying in Global Population for Lipid Homeostasis**, submitted by **Himanshi Garg**, to the Indraprastha Institute of Information and Technology, Delhi, for the award of the degree of **Master of Technology**, is a bona fide record of the research work done by him under our supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Dr. Arjun Ray
Thesis Supervisor
Assistant Professor
Dept. of Computational Biology
IIIT Delhi, 110020

Place: New Delhi

Date: 15th March 2023

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisor, Dr. Arjun Ray, for his guidance, support, and encouragement throughout the course of this research project. Without his expertise and patience, this thesis would not have been possible.

During my thesis work, I used to go to my PhD mentors Gayatri Panda, Sukriti Sacher, Apoorva Mathur and Dr.Sumit Patiyal for my doubt clarifications and I am so thankful for their time, support and guidance. They always tried to come up with new suggestions, that has really helped me shape my thesis work nicely. Their guidance and insights have greatly improved the quality of this work.

Finally, I would like to thank my family for their love and support throughout this process. Their unwavering belief in me has been a constant source of motivation and inspiration. I also thank all the faculty members and staff of the Department of Computational Biology and IIIT Delhi for always helping us throughout our college journey.

I am deeply grateful to all of you for your help and support.

ABSTRACT

Lipid homeostasis refers to the balance of the levels of various lipids, such as fats and cholesterol, within cells and organisms. This balance is crucial for many physiological functions. Disruptions to lipid homeostasis, often caused by missense mutations in proteins involved in lipid metabolism, can lead to conditions like hyperlipidemia, increasing the risk of heart disease and other health problems. Currently, there is significant ongoing research aimed at deciphering the genetic basis of lipid homeostasis in different populations. This research could help develop personalized approaches to managing lipid levels, such as targeted medications or dietary interventions. To contribute to this area of research, we conducted a study focused on identifying common genetic variations associated with the regulation of lipid balance in different populations. Our study involved determining the prevalent mutational patterns and understanding their implications of nsSNPs at the protein structure. The study found unique amino acid exchange patterns among different ethnicities, with low-frequency substitutions more common in the RCT pathway. The African population had a distinct amino acid substitution frequency, and more inter-class conversions were observed when analyzed the impact of nsSNPs on chemical characteristics, indicates a significant impact on protein structure. In the Indian population, the majority of variants were found in the domain region. Structural analysis of 90 variants in 33 genes across six populations revealed that certain non-synonymous variants led to changes in protein secondary structure, caused destabilization, and impaired protein-protein interactions critical for proper lipid metabolism functioning. The study highlights the importance of investigating the effect of variants on protein structure to understand their implications in lipid-related diseases. The findings also suggest that the African and Indian populations may have undergone some genetic divergence leading to distinct amino acid substitutions.

KEYWORDS: Lipid homeostasis ; RCT pathway; inter-class conversions ; genetic divergence ; lipid-related diseases ; non-synonymous variants ; African ; Indian ; CVD

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

The contemporary obesity epidemic is a result of changes in the amount and composition of dietary fat brought about by the industrial revolution [1]. Lipid-composition sensors are employed in lipid biosynthesis to control the activity of lipid enzymes based on the characteristics of lipid composition. To move lipids from one organelle to another, lipid transfer proteins (LTPs) control networks of diffusing lipid droplets or direct exchange at membrane contact points. The lipid composition is also maintained through the transfer of membranes across organelles. The ER is crucial for lipid trafficking as the primary secretory organelle and the exchange network that transports lipids between organelles. Although transfer proteins are quite selective in the lipids they transport, cells can maintain very erratic lipid distributions by carrying out many unidirectional transfers between organelles [2].

Lipids play a variety of roles in membranes, including fundamental structural components, energy storage molecules, signaling molecules, and chemical markers for specific membranes. Since only a small portion of cellular membranes are uniformly dispersed for lipid production, the majority of organelles must rely on lipid transport systems to obtain their complete supply of lipids [3].

To maintain the lipid balance within the human biological system, lipid homeostasis also shares some biological mechanisms with cholesterol efflux, cholesterol-homeostasis, cholesterol transport, endocytosis, excretion, fatty acid oxidation, lipid storage, lipid transport, and reverse-cholesterol-transport [4]. Enzymes such as lipases, hydrolases, and transferases play important roles in these processes. In addition, various hormones and signaling pathways, such as insulin signaling and the hypothalamic-pituitary-adrenal axis, help to regulate lipid homeostasis. Bile acid and pancreatic lipase work together to primarily break down dietary lipids in the small intestine. To make lipids more accessible to lipase, bile acid emulsifies them and serves as a detergent to break up big globules of fat into smaller micelles. Pancreatic lipase converts triglycerides into monoglycerides, free fatty acids, and glycerol. These derivatives enter intestinal epithelial cells called enterocytes and trigger their recombination to become triglycerides. Different lipoproteins with unique surface protein configurations and functionalities are produced by lipoprotein assembly based on their density. Water-insoluble lipids can also be transported by lipoproteins in aquatic conditions. These proteins act as address tags that identify the final location and subsequent function of each lipoprotein [5]. By transferring triglycerides and cholesterol into chylomicrons, which are big lipoprotein particles that leave enterocytes and travel through lymphatic capillaries before entering the bloodstream, chylomicrons carry triglycerides and cholesterol to tissues. An enzyme called LPL can be seen on the blood capillary walls (Lipoprotein Lipase). Triglycerides are turned into fatty acids and glycerol by this enzyme, which then allows them to flow through capillary walls and into tissues where they are either oxidised for more energy or re-esterified for storage. Endogenously generated fats from the liver are packed inside the VLDL lipoprotein and transported to tissues where triglycerides are similarly eliminated [6]. When necessary, hormone-sensitive lipase, which reacts to hormones like epinephrine, mobilizes fat stored in adipose tissue for energy production.

Multiple pathways that are, at least partially, "interdependent" and "cross-regulated" are involved in lipid metabolism. Our research has been primarily concerned with the investigation of various pathways involved in lipid metabolism, with a particular focus on the Reverse Cholesterol Transport pathway. It is well known that there is significant genetic variation among different global populations, with different populations having different frequencies of different genetic variants. This genetic variation can affect many different traits and diseases, including those related to lipid homeostasis. As a result of the discovery of abnormally elevated levels of intermediate metabolites in patients and the demonstration of disease-causing mutations in the genes encoding the implicated enzymes, eight different inherited disorders have currently been associated with various enzyme defects in the cholesterol biosynthetic pathway.

Both de novo production and cellular absorption of low density lipoprotein particles carrying esterified cholesterol are mechanisms by which cells obtain cholesterol [7]. All animal cells contain cholesterol because it is an essential component of the cell membrane, keeps the membrane fluid, has a dynamic and metastable equilibrium, and takes part in a variety of cellular processes. Cholesterol has a critical role in the formation of cell membranes as well as in the promotion of cell development and differentiation

[8]. In the bilayer of the plasma membrane, it is distributed asymmetrically [9]. Bile, steroid hormones, and vitamin D all start as precursors in the presence of cholesterol.

The isoprenoid biosynthetic pathway, which generates a large number of biomolecules known as isoprenoids and contributes to cholesterol synthesis, also plays important roles in signal transduction pathways, cell growth and differentiation, protein glycosylation, and mitochondrial electron transport, among other essential cellular processes. Acetyl-CoA is the first step in the biosynthesis of isoprenoids. With the aid of six additional enzyme activities, acetyl-CoA is transformed into isopentenyl-PP, the fundamental C5 isoprene unit that serves as the building block for all other isoprenoids. The first intermediary exclusively used in the synthesis of sterol isoprenoids, squalene (6 isoprene units), is converted by cyclization into lanosterol (4,4,14-trimethylcholesta-8(9),24-dien-3-ol). At least eight different enzyme reactions are needed to convert lanosterol into cholesterol, including one desaturation between C-5 and C-6, two demethylations at C-14, one isomerization of the double bond at position 8(9) to position 7, three reductions of the double bond at position Δ24 and three reductions of the double bond at position 14 and seven isomerizations of the double bond at position Δ8(9).Since most cholesterogenic enzymes (or enzyme complexes) can handle various sterol intermediates, as is also clear from the sterol intermediates that accumulate in the various biosynthesis defects, the sequence of the enzyme steps involved in the conversion of lanosterol into cholesterol may vary depending on the tissue in which they take place [7].

The formation of cellular cholesterol is controlled by negative feedback. Its production is encouraged by low intracellular cholesterol concentrations while inhibited by high levels. The complex interconnections of sterol intake, absorption, and denovo synthesis, daily losses from enterocyte shedding, and the controlled export of intestinal lipoproteins, predominantly chylomicrons, regulate intestinal cholesterol homeostasis [1].

The precise regulation of cholesterol homeostasis is influenced by a number of feedback systems. The major regulatory targets of these feedback mechanisms are cholesterol 7-hydroxylase in cholesterol catabolism, low-density lipoprotein (LDL) receptor in cholesterol uptake, and 3-hydroxy-3-methylglutaryl coenzyme A (HMG-CoA) reductase in cholesterol synthesis. The liver & intestine are the major sites where cholesterol metabolism takes place. Due to its link to cardiovascular disease risks such as atherosclerosis, in which plaque builds up in the artery walls and reduces or blocks blood flow, cholesterol has gained more significance during the past 50 years. Although the exact origin of atherosclerosis is unknown, many researchers believe that damage to the inner artery wall is the first step in its development. In response, monocytes travel for rescue and transform into arterial cholesterol-loaded macrophages, which engulf the cholesterol molecules and change into lipid-laden macrophages that are a rich source of cholesterol. Over time, substances traveling in the blood, such as cholesterol, and fats begin to build up inside the damaged arterial wall and initiate inflammatory chemical release from the damaged site. As the plague builds up, the arterial wall becomes rigid, grows in size, and the plague is released into the bloodstream, which can stop blood flow and cause clot formation. As a result, the area surrounding the partially blocked artery will receive a limited blood supply, which may cause tissue to deteriorate and even die. This could, in rare instances, result in a heart attack or stroke.

Glomset and Wright proposed a concept of Reverse Cholesterol Transport in general cholesterol efflux from the whole periphery to the liver and its eventual fecal excretion about 50 years ago [10]. Reverse cholesterol transfer is the mechanism through which non-hepatic cells deliver cholesterol to the liver for degradation and then return it through the plasma compartment. The good cholesterol HDL mediates the reverse cholesterol pathway and has been hypothesized to have several anti-atherogenic characteristics. The apolipoprotein reservoir HDL can absorb unesterified cholesterol. ApoA and ApoC are important parts of HDL, but ApoA1 is what distinguishes it from other particles. The intestine and liver both can release HDL. While other particles are more triglyceride or cholesterol-rich, HDL is more protein-rich. With its anti-atherogenic action, HDL helps to maintain the net cholesterol balance in the arterial space by removing excess deposition of fats from the artery space and recycling it in the liver. HDL particles must pass through the extra-hepatic space and interact with the ATP-binding cassette transporter A1 (ABCA1), which is present in nearly all peripheral cells, in order to remove excess cholesterol from the extra-hepatic space and convert it to cholesterol ester with the help of the LCAT enzyme. The liver has a multifunctional receptor called the scavenger receptor class B type I (SR-BI). It mediates bidirectional lipid transport in macrophages, which is dependent on the amount of cholesterol in lipid-laden macrophages. It scavenges the HDL particle, absorbs all the cholesterol inside it, and then releases the HDL particle for further scavenging. With the aid of ABCA1 and ABCG1 receptors, the nascent HDL particle absorbs cholesterol; as its size grows as a result of cholesterol deposition, the nascent HDL particle develops into an HDL3 particle. HDL3 particles would change into HDL2 particles as a result of having higher cholesterol, triglycerides, etc. Last but not least, the liver receives these HDL2 particles for recycling. IDL and VLDL give triglycerides, but HDL can interact with these particles and furnish them with proteins. Eventually, these particles will transform into LDL after absorbing the triglycerides from the IDL and VLDL. The liver can also absorb LDL; to achieve this, hepatocyte cells that have the LDL receptor must undergo receptor-mediated endocytosis, which ultimately results in the uptake of cholesterol. In essence, HDL facilitates cholesterol uptake through two different pathways: 1) Indirectly through LDL particles, and 2) Directly through binding to the scavenger receptor. The efficacy of the "Reverse Cholesterol Transport" (RCT) mechanism has been examined using measurements of the cholesterol concentration in HDL particles, however macrophage-derived cholesterol only makes up a small portion of the cholesterol delivered by HDL particles.

Lipidoses are caused by the inhibition of cholesterol efflux and are a group of inherited metabolic disorders caused by the deposition of fat-like substances to a certain level in the body's cells, organs, and tissues, which is known to cause diseases like CAD (Coronary Artery Disease). People with these illnesses may either be unable to create the lytic enzymes needed to catabolize lipids or they may make the enzymes but they won't function. Particularly in the brain, liver, and spleen, excessive fat deposition over time will result in lasting cellular or tissue damage [11].

The influence of mutation may be related to the disturbance in cholesterol metabolism. While some SNPs occurs in coding regions, certain SNPs take place in non-coding regions. SNPs that occur in non-coding regions of the gene and have no effect on the gene product are known as silent mutations, which imply that the protein sequence remains unaltered. Silent mutations could potentially be found in the coding region because each amino acid is encoded by more than one codon. Non-synonymous SNPs (nsSNPs), which affect the protein sequence, fall on the opposite end of the spectrum [12].

The point mutations that take place in cells are the sole foundation of the scientific theory of evolution. The idea explains the origins and diversity of the Earth's living things. The advantageous changes can aid in reproduction, allowing the positively impacted genes to be passed on to the next generation. Evolution would not be possible without mutation, and mutation would not be possible without any change caused by several factors, such as environmental influences (where DNA interacts with the environment, sometimes to the detriment of genetic information), or replication errors brought on by the incorporation of an incorrect base during the replication. As a cell divides, the DNA duplicates itself. Sometimes, before a wrongly positioned nucleotide is fixed in place and a mismatch is produced in the complementary strand, the enormous network of DNA repair machinery in the cell blocks cell division. The mutation is nonetheless persistent in the cell even if the repair mechanism detects the mistake before the complementary strand develops [13]. Not all SNPs are identical by descent because SNPs might arise from replication faults. Additionally, each Alu insertion adds a structural feature of about 300 bp, making them potentially more likely than SNPs to have an impact on daily life[14].

Another theory is that DNA deterioration results from exposure to specific chemicals or radiation. DNA becomes altered as a result of the thymine dimers being broken by UV light. Scientists have discovered a very high frequency of this UV-induced mutation in the genes linked to basal cell carcinoma, a type of skin cancer. All these mutations lead to evolution[15]. Natural selection causes this to happen even if the mutation rate in mammalian cells is quite modest. Harmful mutations can harm an organism's capacity to reproduce or even cause it to perish[16]. The main driving forces behind genetic diversity among populations are mutation, genetic mixing, genetic drift, and natural selection.

The ratio of retained to repaired DNA damage is insufficient for those experiencing the terrible effects of the unrepaired mutation. However, this ratio is crucial for evolution[**?**].

The figure 1.1 represents the missense mutation, which happens when a single nucleotide base in a DNA sequence is exchanged out for another one, has an impact on the fundamental dogma. This mutation results in a new codon and, as a result, a different amino acid. These mutations are quite frequent and often have little impact on the protein's overall structure or activity. However, when two amino acids with very dif-

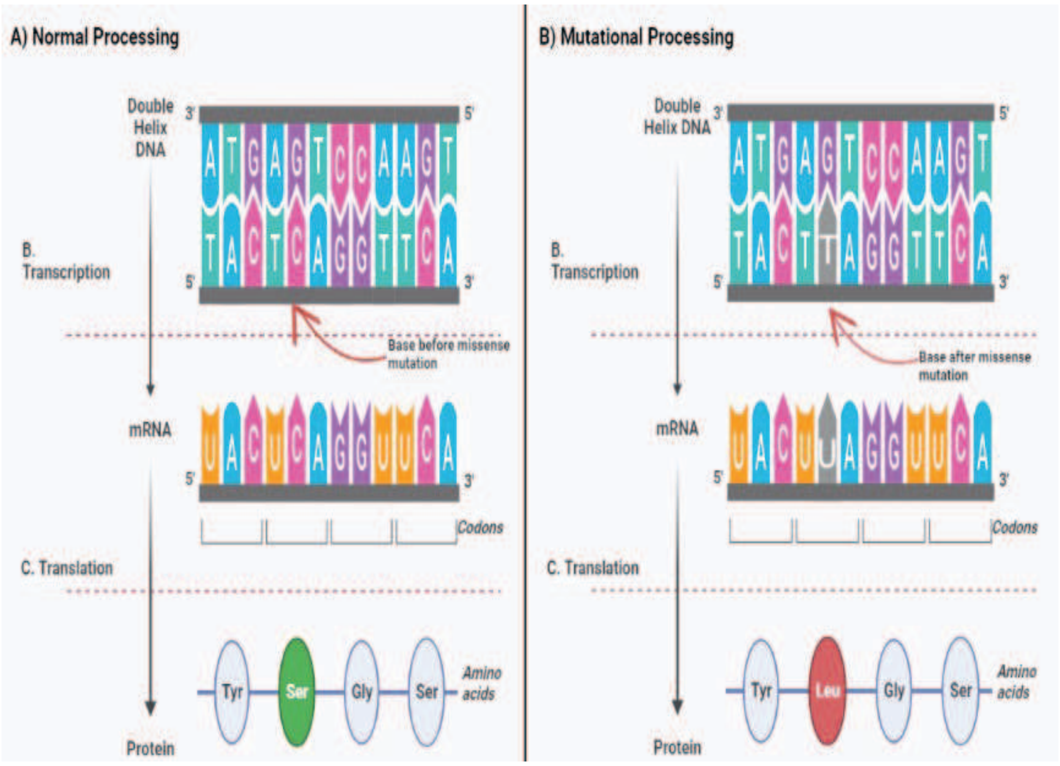


Figure 1.1: Mutational Effect over Central Dogma

ferent chemical properties are substituted, missense mutations can drastically change how a protein behaves. In such cases, they are referred to as non-conservative missense mutations.

The majority of changes to protein structure and function are neutral. Therefore, the likelihood that they will be replaced increases with how similar the amino acids are[17]. The mutational effect is pleiotropic, and a mutation that favorably affects a protein's activity may adversely influence the protein's stability and lower the amount of soluble, functional protein. The future function of a mutation that is advantageous for an alternative can also be harmful or neutral for the protein's current function [18]. The presence of an SNV in a gene that is functionally significant has a higher likelihood of having a negative effect by either affecting the regulatory mechanism or by causing structural changes at the protein level that result in disruption, even though only non-synonymous mutations in DNA can result in amino acid replacements and that the evolution of proteins is slower than that of DNA [19].

Single nucleotide polymorphisms provide insight into the genetic basis of complicated human disorders and phenotypic variations [20]. Rare mutations hardly ever have a significant impact on how genetic diversity has developed in human populations [21].

SNPs are considered single nucleotide polymorphisms if they are found in more than 1% of the population (SNPs). Despite the fact that missense mutations are implicated in more than half of all known inherited diseases, their molecular consequences are typically less dramatic than those of mutations that modify the mature peptide more significantly. Missense mutations can change a protein's structure or function in at least two separate ways: either they destabilize the entire protein fold, or they change the functional residues that control active sites or protein-protein interactions. At protein interfaces as well as in proteins' hidden areas, pathogenic mutations are more common [22]. In terms of their impact on lipid homeostasis, missense mutations can affect the function of proteins involved in lipid metabolism, leading to abnormal levels of lipids in the body. This can result in conditions such as hyperlipidemia (elevated levels of lipids in the blood) or hypolipidemia (low levels of lipids in the blood), which can increase the risk of heart disease and other health problems. Therefore, it is crucial to interpret the number of mutations and how they impact the protein's structure, stability, and function [23].

When each participant in the study is identified according to their country of origin, we can see that some variations are more prevalent in one region of the world than in another [24]. A population's genetic makeup, migration patterns, and evolutionary history can all be inferred from the distribution of SNV allele frequencies. People with different geographic ancestries exhibit genetic variances with differing frequencies of carrying any SNP. Genetic diversity is one of the elements that contribute to the inter-individual disparities among different ethnicities [19]. As a result, individual likelihood of carrying an allele frequency distribution changes depending on their geographic ancestry. Researchers can find out the allele frequency of a variant and predict the impact of that variant based on its rarity or commonness because detrimental alleles are often assumed to display lower frequencies in a population than benign alleles [25]. The majority of genetic variant data originates from databases like the 1000 Genome and gnomAD, which provide data on ethnicity-based variants that are predominantly biased toward Eurocentric [19] implying a severe ancestral bias problem in human variant sequencing efforts. As a result, most populations of different ethnicities are not subject to allele frequency queries by researchers, which suggests that there may also be population-specific mutations in underrepresented communities. As a result, this emphasises the importance of comprehensively analyzing how SNVs affect people around the world.

To study the influence of missense mutations on lipid homeostasis, for this research the 1000 Genomes Project sample data was used, which includes genomic information. By analyzing this data, one can identify missense mutations in genes involved in lipid metabolism and compare their frequencies in individuals with and without abnormal lipid levels.A complete set of synonymous (sSNPs), indels (insertions and deletions), and non-synonymous (nsSNPs) modifications for 1092 people from various populations have been made public by the 1000 Genomes Project. According to the 1 kG data, each person will typically differ from the reference human genome sequence at (10,000–12,000) synonymous and (10,000–11,000) non-synonymous places. It can be used to look at the differences and structural preferences between known benign mutations and known disease-associated mutations, or at the very least between known benign mutations and known disease-associated mutations [17].

The IndiGen effort was started with the intention of gathering sequencing data from thousands of people in India representing various ethnic groups and creating public health technology applications using this population genomic data [19].

In order to determine the impact of population-specific variations producing a difference due to different ethnicity, we conducted an extensive and comparative analysis of common variants among varied populations (using 1000G and IndiGen data) in our current work. This study focuses on examining many elements of how non-synonymous mutations affect the stability, functionality, and structure of proteins. According to study, it has been discovered that a variety of disorders are brought on by disturbances in the metabolism of cholesterol. These mutations may cause the substitution of the wild-type amino acid with the mutant type, which could have a phenotypic effect or not. Some of these diseases include Tangier Disease, which develops in the ABCA1 gene as a result of the replacement of wild-type amino acid residue with an alternate amino acid at various locations in the protein sequence, and Hyperlipoproteinemia-3, a disease that develops as a result of variation in the APOE gene and many more. Therefore, it's critical to evaluate the amount of mutations and how they affect the protein's structure, stability, and function. Any protein destabilisation brought on by non-synonymous SNV (nsSNV) might alter the routes, intermediates, or end product, and will eventually cause a disruption in the homeostasis of the pathways within the human biological network.

To decipher the genetic bias underlying global population variation for genes involved in lipid homeostasis, the genome-wide association study (GWAS) is needed to conduct. This involves comparing the genomes of individuals from six different global populations to identify genetic variants that are associated with lipid homeostasis. The identified variants can then be compared across populations to identify any differences in their frequencies. In order to investigate the impact of missense mutations on pathways and the structural alterations brought on by these mutations at the protein level, this study was conducted at both the sequence and structural level. By assisting us in comprehending the variability brought on by these variants, this work may help us figure out how SNV affects the effectiveness of the pathways.

CHAPTER 2

MATERIALS & METHODOLOGY

2.1 Data Collection

The 1000 Genomes Project (1000GP) aims to provide a comprehensive understanding of human genetic variation through the sequencing of numerous individuals. This collection includes about 2,500 genomes from 26 populations that were sequenced during the three phases of the project, and it is useful for evolutionary, functional, and medicinal research in human genetics by providing a list of variations and haplotypes that can be utilized. The individuals were divided into five major continental groups: East Asia (EAS), Admixed America (AMR), Africa (AFR), Europe (EUR), and South Asia (SAS) [26].

To investigate this genetic variation for the individuals from diverse ethnicity's, we retrieved Variant Call Format (VCF) files from the GRCh38 assembly that had been processed by the 1000 Genomes Project (1KG) from the International Genome Sample Resource (IGSR), which maintains and shares the 1000 Genomes Project's sample data on human genetic variation. Moreover, the project's dataset provides a natural background for human amino acid germline mutations, making it an appropriate choice for our analysis [17]. By leveraging this dataset, we were able to gain insights into the genetic variation present in different populations and identify potential targets for further investigation.

India accounts for 16.7% of the global population with a total of 1.21 billion inhabitants. Due to a complex demographic history, Indian populations are highly diversified, consisting of thousands of endogamous sub-populations with varying levels of mixing and socioeconomic structure, representing multiple ethnic and linguistic lineages. The four main linguistic lineages are Indo-European (IE), Dravidian (DR), Tibeto-Burman (TB), and Austro-Asiatic (AA) [14], but these lineages are underrepresented in global sequencing datasets. The bulk of data pertaining to genetic variants is sourced from databases such as the 1000 Genome database [27] and gnomAD [28], which contain information on variants categorized by ethnicity, primarily with a Eurocentric bias. This can be attributed to the fact that most studies aimed at identifying genetic variants associated with diseases, such as Genome-Wide Association Studies (GWAS), have been conducted mainly on individuals of European ancestry (78%), followed by Asian (10%), African (2%), Hispanic (1%), and other ethnicities (<1%) [29], thus underserving the Indian population. In an effort to improve public health applications, the Council for Scientific and Industrial Research (CSIR) with funding from the Government of India, initiated the IndiGen Genome Project, which aims to collect and sequence the genomes of ethnic Indian people for predictive medicine using repeats and single nucleotide polymorphisms.[30]. The Indian Genome Variation (IGV) consortium study highlighted the diversity of the Indian population through SNP-based genotyping of 900 genes in over 1800 individuals from 55 sub-populations. The study expanded knowledge of precise genetic markers defining various genotype-phenotype associations and identified novel founder mutations specific to the Indian subcontinent [31]. Hence for this comparative analysis, the variant information with allele frequencies for the Indian population was obtained from the IndiGenomes database.

Both the projects (1kG and IndiGen) involved characterizing biallelic and multiallelic SNPs, indels, and structural variants. These variants were annotated in accordance with the GRCh38 human reference genome using the ANNOVAR command-line tool based on Perl [32]. Additionally, the annotated variants were filtered out specifically for non-synonymous mutations from the ANNOVAR processed output which were considered for analyzing their implications over the protein structure and function.

2.1.1 Lipid Homeostasis Genes Collection

The distribution of lipids within and between cell membranes is irregular, defining the identity of each organelle. Local metabolic pathways and pathways that move lipids are necessary for this distribution, and feedback mechanisms regulate lipid levels and fluxes to decode topographical information in organelle membranes [33]. Complex homeostatic processes are crucial for maintaining cellular and organismic homeostasis by regulating membrane characteristics and functions [2]. Lipid homeostasis and inflammation are critical factors in complex disorders like atherogenesis, which requires lipid-rich foam cell macrophages for the development of atherosclerotic lesions.

Obesity-induced cardiomyopathy has disrupted cellular lipid homeostasis, which must be maintained to prevent such disorders [34]. Although previous analyses found many genes essential for maintaining lipid homeostasis, a list of 182 genes was created by annotating key characteristics such as Transcript Start-Transcript End, Ensembl IDs, PDB IDs, RefSeq Match Transcript, Gene Start-Gene End, and Uniprot ID using the BioMart resource [35].

2.2 Computational Requirements

This section describes the system requirements and setup used for the investigation. Data processing was carried out at IIIT Delhi using the open-source, Unix-based Linux operating system. Although we preprocessed the data on Linux, the scripts were also compatible with Windows when a Python compiler was used. To set up the Python environment, we required a computer system with administrative access, Linux operating systems, and an internet connection. These prerequisites were provided by RayLab at IIIT Delhi. Python is a simple and flexible programming language, named after the British comedy troupe Monty Python, and first released in 1991. The developers aimed to create an entertaining programming language. The latest version of the language, Python 3, is believed to have a secure future. After setting up the workstations, we installed and configured Python using the command line. We ensured that the Python version was up-to-date by upgrading the system using apt-get. Once the upgrade was complete, we checked whether Python3 was installed on the machine or not. We then installed all the necessary Python packages or libraries, such as Pandas and Numpy, using pip or the conda package. In addition, Linux offers the option to set up a Virtual Environment, a user-defined sub-environment within the Linux workspace. This ensures that each project has its own set of dependencies that won't conflict with any other projects. Virtual environments allow for greater control over projects and the ability to use multiple versions of packages on the same workstation. We can create as many virtual environments as needed by first installing the venv module, which is part of the default Python 3 library, and then choosing a directory to create the Python environment. After generating the environment and installing all the requirements, the virtual environment can be activated using the activation command that calls up the activate script.

2.2.1 Jupyter Notebook Installation

Jupyter Notebook is a web-based tool that is open-source and allows users to create and share documents containing live code, equations, visuals, and text. It is commonly used for data analysis, scientific computing, and machine learning tasks.

With Jupyter Notebook, you can write and execute code, visualize data, and document your work all in one place. The notebooks are saved in a file format with the .ipynb extension.

Jupyter Notebook is a preferred tool for researchers because it provides a convenient and powerful way to analyze and visualize data, collaborate with others, and document their work. It is also extensively used in education, as instructors can create interactive, self-contained lessons that can be easily shared with students.

To utilize Jupyter Notebook, it must first be installed on your computer using the pip command. This will also install all required dependencies for the tool. Once installed, the notebook can be launched from the terminal.

After launching Jupyter Notebook, you can create a new notebook by selecting the "New" button and specifying the desired kernel (e.g., Python, R). Then, you can enter and execute code in cells, add markdown text to document your work, and use various tools to visualize and analyze data.

2.2.2 Setting up python environment

Some common options for setting up a Python environment on your computer:

1) Install a standalone Python distribution: If you don't need any specific libraries or) frameworks and want to use Python for general-purpose programming, you can download and install a Python distribution such as Anaconda or Python.org Python. These distributions come with the Python interpreter, a package manager (e.g., pip), and a set of basic libraries.

2) Use a virtual environment: A virtual environment is a self-contained Python environment that allows you to install packages and libraries specific to a particular project. This is helpful if you need to use different versions of a library for different

projects or want to have separate environments for different projects. You can create a virtual environment using the "venv" module in Python or a tool like "virtualenv".

3) Use a containerization platform: You can package your application and its dependencies into a container using containerization platforms such as Docker. This allows you to run your application on any machine with the containerization platform installed and share it with others.

In this project, we used Pandas, NumPy, matplotlib, seaborn for data visualization and pre-processing, and Bio-python for data fetching from NCBI. These libraries were installed in the same environment using different commands.

2.3 Tools used for modeling and screening

2.3.1 ANNOVAR

ANNOVAR is a Perl-based application that can run on a wide range of hardware platforms with standard Perl modules installed. To install Perl, we can use commands such as sudo, apt-get, or pip similar to installing Python. Once Perl is installed, we can download all the databases required for annotation. ANNOVAR is a computationally intensive task that can overload local machines, so we used HPC instead of local machines for annotation. For this particular run, we used an HPC server with the following specifications: - RAM: 512 GB - CPU: 32 - Cores per socket: 8 - Duration: 48 hours.

2.3.2 Pfam HMM

PfamScan is a tool that can search a FASTA sequence against a library of Pfam HMM. It can be accessed via the web, REST API, GUI or as a standalone tool. Our objective was to identify the protein domains and predict the location of interactive residues given a protein sequence. To achieve this, we utilized the web form of PfamScan. The process involves multiple steps, where the first step is providing user input, such as the sequence or database. In the second step, we modified the default tool parameters to suit our needs. Finally, we submitted the analysis with a name or title attached, by clicking the submit button.

2.3.3 Modeller

Protein homology modelling, also referred to as comparative modelling, utilizes a standalone software program named Modeller. The homology modelling technique is based on the observation that the tertiary structure of a protein is more conserved than its amino acid sequence. Proteins that share significant structural similarity despite having divergent sequences are expected to have comparable structural characteristics. Since obtaining experimental structures for every protein of interest is difficult, expensive, and time-consuming, techniques such as X-ray crystallography and protein NMR are often not feasible. Homology modelling can provide valuable structural models that aid in the development of hypotheses about a protein's structure and function. Research has demonstrated that naturally occurring homologous proteins and evolutionary related proteins share similar sequences and protein structures. Moreover, a 3D protein structure is evolutionarily more conserved than expected solely based on sequence conservation. In homology modelling, the quality of the model is determined by the quality of the sequence alignment and template structure. As the output model quality increases, sequence identity decreases. The two primary causes of major inaccuracies or faults in homology modelling are low sequence identity or errors in the sequence alignment and template selection. The modelling process can be broken down into four steps: template selection, target-template sequence alignment, model development, and model evaluation.

2.3.4 Gromacs

GROMACS (GROningen MAchine for Chemical Simulations) is a widely used open-source software for molecular dynamics simulations. Its applications span across various scientific fields such as chemistry, biochemistry, and physics to simulate the behavior of biological and chemical systems, including proteins, lipids, and nanoparticles. GROMACS employs classical molecular dynamics principles to simulate the motion of atoms and molecules based on the concepts of classical mechanics. It can simulate a diverse range of systems, from simple biological to complex chemical systems, with varying levels of precision and detail. Additionally, it is useful in analyzing the kinetic and thermodynamic properties of molecules and optimizing their geometric structure.

GROMACS is coded in C and C++ and is compatible with Windows, Linux, and macOS. It is freely available under the GNU General Public License and can be obtained from the official GROMACS website.

2.3.5 Modrefiner

We utilized modrefiner to enhance the model files generated by Modeller. Modrefiner offers a refining algorithm that operates at an atomic-level, suitable for high-resolution protein structures. During the refining process, the side chains and backbone atoms are completely flexible, while the conformational search is controlled by a combination of physics and force field information. The aim of this approach is to bring the hydrogen bonds, backbone structure, and side-chain placement of the basic models closer to their native state. Furthermore, modrefiner also improves the physical integrity of adjacent structures.

2.3.6 DSSP

The DSSP method predicts the secondary structure of proteins based on their three-dimensional (3D) structure. Secondary structure refers to the regular folding patterns in a protein's polypeptide chain, such as alpha helix, beta sheet, and turns.

DSSP is based on the discovery that the arrangement of hydrogen bonds between a protein's main chain amide groups greatly influences its secondary structure. Using these hydrogen bonds, DSSP categorizes each amino acid residue into one of eight potential secondary structure classes namely, Alpha helix (H), Beta bridge (B), Extended strand, or beta sheet (E), Bend (S), Turn (T), Coil (C), 3/10 helix (G), Pi helix (I).

The GROMACS software package includes DSSP as a command-line program. It can be used to analyze molecular dynamics simulation results or predict a protein's secondary structure based on its 3D structure in a PDB file.

2.3.7 Dynamut2 API

The web server known as DYNAMUT2 applies a combination of machine learning and physics-based models to predict and analyze the impact of mutations on protein stability and function. It enables users to submit a protein sequence and a list of mutations for analysis, generating a graphical representation of the predicted effects and forecasting how each mutation would affect the protein's stability and functional capacity. The web server also provides comprehensive information about the underlying models and assumptions used to generate the predictions.

2.3.8 HBPlus

HBplus is a software tool that predicts and analyzes hydrogen bonds in biomolecules including proteins. The software operates on the principle that hydrogen bonds are crucial for determining the 3D structure and function of biomolecules. HBplus combines geometric and energetic criteria to identify hydrogen bonds in a protein structure. The software is capable of predicting both main chain and side chain hydrogen bonds and can analyze their patterns and characteristics. It is a command-line utility that can be used on Windows, Linux, and macOS to anticipate and analyze hydrogen bonds in a protein structure in a PDB file, or to evaluate molecular dynamics simulation outcomes.

2.4 Data Preparation

2.4.1 Sequence Analysis

Variations in the fundamental structure of proteins caused by incorrect amino acid substitutions are referred to as sequence variants. Detecting these variants early is crucial in product and process development because they can result in protein misfolding and aggregation, which can ultimately affect therapeutic safety and efficacy. There are two types of sequence variants based on their source: mutations and misincorporations. While amino acid misincorporation can only be detected at the protein level, mutations are DNA-level defects that typically arise from the accidental integration of incorrect nucleotides during DNA replication [36]. To analyze single nucleotide variations in various populations, sequence-level analysis was performed based on amino acid exchanges. The data used for the sequencing analysis included genes involved in maintaining lipid balance within the human biological system and the result of the variant calling file processed by ANNOVAR.

2.4.2 Structure Analysis

The omics revolution has led to a significant increase in the number of human mutations associated with disease. Many of these mutations occur in protein-coding regions of the genome and can affect the structure and function of the protein, ultimately impacting the phenotype. Understanding these structural and functional impacts can aid in the design of future studies, potentially leading to the development of more accurate diagnostic tools and therapeutic drugs[37].

To investigate the effects of point mutations on protein structure and their negative impacts, a structural analysis of mutants was performed. Fasta files for the study were obtained from NCBI Genbank using a Python script for the lipid homeostasis genes, and variant sequences/models were created by altering the native residues of the protein sequence. Various structural analysis techniques were then used to predict how these variations would affect the protein's structural characteristics, including hydrophobicity, solvent accessible area, and hydrogen-bond network calculations.

Before undergoing structural-level analysis, the raw sequence data was filtered using several necessary filters. These filters included:

1. Accessibility of protein crystal structures,

2. Protein crystal structures with;

$$SequenceCoverage \geq 70\%$$

, 3. Non-synonymous SNVs observed in the populations with;

$$AlleleFrequency \geq 10\%$$

. After applying these filters, 79 genes with their corresponding 175 variants for the

AFR population, 69 genes with their corresponding 140 variants for the AMR population, 70 genes with their corresponding 132 variants for the EAS population, 68 genes with their corresponding 143 variants for EUR population, 73 genes with their corresponding 140 variants for the SAS population and 34 genes with their 67 corresponding variants for the Indian population were filtered out. Applying the two most crucial filters to the output of the filtered data from above will allow for further investigation:

1. Chopping of the signal peptidase.

2. Availability of the correct native residue within the protein chain at the exact position in Isoform-1.

The input data for the structural data outlined like for the AFR population, there were 20 genes with associated 32 variants; for the AMR population, there were 15 genes with associated 23 variants; for the EAS population, there were 17 genes with associated 25 variants; for the EUR population, there were 16 genes with associated 23 variants; for the SAS population, there were 15 genes with associated 23 variants; and finally, there were 29 genes with associated 53 variants for the Indian population.

Several databases, including the Human Gene Mutation Database (HGMD), the Online database of Mendelian Inheritance in Man (OMIM), and the UniProtKB Human Polymorphisms & Disease Mutations collection (HumsaVar), gather data on inherited disorders linked to variants. However, HUMSAVAR data was employed for doing a comparative structural analysis of the genes with their matching and previously known disease-associated variations. Disease-related variations from the literature and OMIM are present in the Humsavar database. As of December 2011, OMIM (Online database of Mendelian Inheritance in Man) reported on about 10,200 nsSNPs linked to diseases, while Humsavar reported on about 23,500 nsSNPs linked to diseases[17].

2.5 Data Processing

2.5.1 Data Annotation

ANNOVAR is a bioinformatics tool that efficiently annotates genetic variants from high-throughput sequencing data, predicts their functions, and prioritizes SNVs, indels, and CNVs in a particular genome. It uses the genomes of model species like humans, mice, and others, as well as human reference genomes hg18, hg19, and hg38 for annotation. ANNOVAR sets the standard for defining genetic loci and accepts text-based data files in vcf and bed formats. It modifies the vcf file, annotates it, and outputs an Excel-compatible file.

ANNOVAR supports three types of annotations: gene-based, filter-based, and region-based. Gene-based predicts the functional effects of gene variations, filter-based provides extensive information on the variant, including population frequency and deleteriousness prediction scores, and region-based predicts the variants' association with specific genome regions. For the study, the filter-based annotation type was chosen as it works on mutations and compares variants with things like the $A- > G$ alteration at position chr1:1000-1000 [38].

2.5.2 Amino acid Exchanges

A single-point mutation can occur in the genome's sequence when a single nucleotide is substituted, resulting in missense mutations induced by nsSNPs. These mutations involve the substitution of an amino acid in the protein sequence and can either be polymorphic missense mutations or rare missense mutations that only affect a small group of individuals, such as a family, depending on their prevalence [39]. An amino acid substitution can disrupt a critical folding site, preventing the formation of the folding nucleus and causing the rest of the structure to rapidly condense [40]. Analyzing amino acid substitutions is a valuable approach to investigate changes in a protein's physico-chemical properties and identify patterns in the impact of variants at a physico-chemical level. To facilitate this analysis, a Python script was developed and employed to investigate reported variants in genes related to lipid balance. Specifically, the script evaluated the frequency of amino acid exchanges in each population, enabling the rapid assessment of each population's pattern of amino acid exchanges. The script generates a matrix by normalizing the data, such that the values range between 0-1 for each population.

$$Normalization = \frac{Ref - AltFreq}{RefFreq} * \sum \frac{ProbibilisticAlleleFrequency}{Ref - AltFreq(populationspecific)}$$

By investigating the chemical changes caused by genetic variations in different populations, one can identify the most common classes of amino acid substitutions. Once amino acid substitutions were identified, they were classified into different classes based on the chemical properties of the amino acid involved. For example, substitutions involving amino acids with similar properties, such as charged or hydrophobic amino acids, were classified into the same class.

2.5.3 Statistical Analysis

Statistical analysis involves organizing and interpreting data using established mathematical rules and procedures to form a statistical hypothesis. In this study, the occurrence of amino acid pair exchanges in IndiGen data and 1000G population data (AFR, AMR, EAS, EUR, & SAS) was compared using the non-parametric Mann-Whitney U Test (also known as the Wilcoxon rank-sum test). Before conducting the test, it was necessary to establish both an alternative hypothesis and a null hypothesis, which is standard practice for statistical analyses. The alternative hypothesis for this study was "There is a significant difference between two populations," while the null hypothesis was "There is no significant difference between two populations." Statistical significance was defined as p-values less than 0.05.

2.5.4 Mutability Analysis

Mutability analysis is essential for identifying which amino acids are more likely to undergo mutation in different populations. This information can be used to understand the genetic diversity and evolutionary patterns of different populations, as well as the potential impact of mutations on protein structure and function. By determining which amino acids are more prone to mutation in specific populations, it could help to develop targeted approaches for disease diagnosis, treatment, and prevention.

Additionally, mutability analysis can help in predicting the impact of amino acid substitutions on protein stability, activity, and interactions. This information would be crucial for studying the molecular mechanisms underlying genetic diseases and identifying potential drug targets. Overall, mutability analysis provides a valuable insight for investigating the functional consequences of genetic variations and their role in shaping the genetic diversity of different populations. The results of the mutability analysis can be visualized using graphical representations such as heat maps or scatter plots. These visualizations can help to identify patterns and trends in the data.

2.5.5 Protein Domain Analysis

Domains are the building blocks of molecular evolution, as they can undergo recombination events to create proteins with distinct functions. Many polypeptides, consisting of several hundred amino acid residues, fold into stable, globular domains. These domains often contain organized secondary structures that form unique functional motifs or domain folds. Different domains serve various roles, such as interacting with other proteins. While small proteins usually have one domain, eukaryotic proteins frequently have multiple domains joined by flexible linkers. Even when removed from the whole protein, domains maintain their 3-D structure and function.

This study aims to understand how single nucleotide polymorphisms (SNPs) in or near domain regions can impact metabolic or signaling pathways, as changes to protein structure and function can have harmful effects on human health. For example, certain genetic variations can cause misfolded or dysfunctional proteins, leading to diseases like cystic fibrosis or sickle cell anemia. By comprehending how these variations affect protein domains, researchers may identify novel therapeutic targets for these and other disorders. Investigating the effects of protein domain variations is also crucial in understanding the function of these domains in cellular processes. Proteins often have specialized roles, such as catalyzing chemical reactions or binding to molecules. Analyzing how variations impact domain structure and function can provide insights into how these processes are regulated and contribute to overall protein function. Mutant residues in the protein domain region would have a more significant impact on protein structure and activity than those found in the pre-domain or post-domain regions.

To create mutants, the multi-sequence alignment (MSA) was performed using Clustal Omega, and a Python script was used to fetch sequence data in FASTA format from NCBI Genbank. Native sequences were modified at the necessary position using ANNOVAR processing. The Pfam HMM library, which is maintained by EMBL-EBI and used for protein domain research, was searched for mutant sequences using the Pfam-

Scan web server [41]. The input file contained all protein sequences in FASTA format and was given default settings. The output file provided information on the location of sequence start and end, HMM(domain name) name, and HMM(domain) start and end, as well as other necessary data.

2.5.6 Variant Model Generation with Energy Minimization & Protein Structure Refinement

The maintenance of Modeller, a software for homology modelling, has been taken over by Andrej Sali from the University of California, San Francisco. The software utilizes an autonomous script that requires only a PDB file as input and automatically carries out sequence alignment, loop refinement, and other modelling processes. Modeller generates 3-D variant models, and to create a probability density function for each atom's location, it uses a technique called "fulfilment of spatial limitations," inspired by protein NMR. This method relies on an input sequence alignment between a template protein with a solved structure in PDB format and the target amino acid sequence to be modeled [42].

To determine the minimal energy conformation for provided structures, Gromacs (Groningen Machine for Chemical Simulations) was employed. Gromacs is a molecular dynamics tool primarily used for simulating proteins, lipids, and nucleic acids. It was first developed in the Biophysical Chemistry department at the University of Groningen and is currently being updated by contributors from universities and research institutions worldwide. The software includes a script that converts PDB files' molecular coordinates into the necessary format. After producing a setup file to simulate multiple molecules, the simulation run, which can be time-consuming, generates a trajectory file that details the atoms' movements through time. This file can be examined or viewed using the various tools available [43].

ModRefiner was used for high-resolution protein structural refinement. The method allows for a full atomic model, a C-alpha trace, or a main-chain model as starting points. During structural refinement simulations, both side-chain and backbone atoms are entirely flexible when the conformational search is managed by a combination of physics and knowledge-based force fields. ModRefiner offers the option of assigning a second structure as a reference for refinement simulations, in addition to bringing the hydrogen bonds, backbone topology, and side-chain positioning of the initial models closer to their natural state. It also significantly improves the physical state of neighboring structures. The standalone application also offers an ab initio full-atomic relaxation feature, in which the refined model is not limited by the initial or reference models [44].

2.5.7 Variants availability in Secondary Structure of the Proteins and estimating their Solvent Accessible Area

The density of interactions between residues is a crucial structural element that determines a protein's ability to tolerate mutations without disrupting its fold. A higher contact density in a particular fold corresponds to greater resistance to mutations [22]. To standardize the assignment of secondary structures, Chris Sander and Wolfgang Kabsch created the DSSP tool. The DSSP database provides secondary structure assignments for all proteins in the Protein Data Bank (PDB) based on the most probable assignments derived from the protein's 3D structure, rather than by expectations [45]. Simon Hubbard and Janet Thornton developed Naccess, a program that predicts the absolute and relative surface accessibilities of protein structures. It can be used to determine the availability of mutant residues in the protein structure, whether on the surface or in the core, and to verify their accessibility in the secondary structure type [46].

2.5.8 Influence of mutation over Hydrogen Bond Network

Proteins, the building blocks of cell membranes along with lipids, are connected by peptide bonds. The stability of proteins in the cellular environment is provided by the bonds that connect the polypeptides to one another, such as vanderWaals force and hydrogen bonds. To determine the hydrogen bonds in a protein, the computer application HBPLUS can be used. The software takes the protein's three-dimensional structure as input and calculates all hydrogen bond geometries, including a list of neighbouring interactions and hydrogen positions. It also offers comprehensive customisation options for H-bond criteria, donor and acceptor atom types, and outputs PDB files. Using the HBPLUS software, H-bond breakage caused by mutations, which can affect protein structure, was analysed. The PDB file included projected polar hydrogen positions [47].

2.5.9 Mutational effect over Protein Stability

Protein stability refers to the net balance of forces that determine whether a protein will maintain its folded structure or be in an unfolded state. Understanding the fundamental thermodynamics of protein folding is one of the many benefits of protein stability [48]. DynaMut2 is a command-line API and web server that utilizes two different normal mode methodologies to analyze and visualize protein dynamics by sampling conformations. It also quantifies the impact of mutations on protein dynamics and stability, which can arise from changes in vibrational entropy [49].

There are three distinct applications of DynaMut2:

1) Calculation $\Delta\Delta$G for single point mutations.

2) Evaluation $\Delta\Delta$G for multiple point mutations (up to three).

3) Analysis of protein dynamics using NMA.

To use the Single Mutation option in DynaMut2, users need to provide a string consisting of the wild-type residue one-letter code, the residue position, and the mutant residue one-letter code. Additionally, users must submit a protein structure in PDB format or a four-digit PDB entry code and specify the chain identifier of the affected chain. If users want to use the List of Mutations option, they need to provide a file containing a list of variants, with each variant conforming to the same mutation code as the Single Mutation option [23].

CHAPTER 3

RESULTS & DISCUSSION

3.1 Sequence Analysis

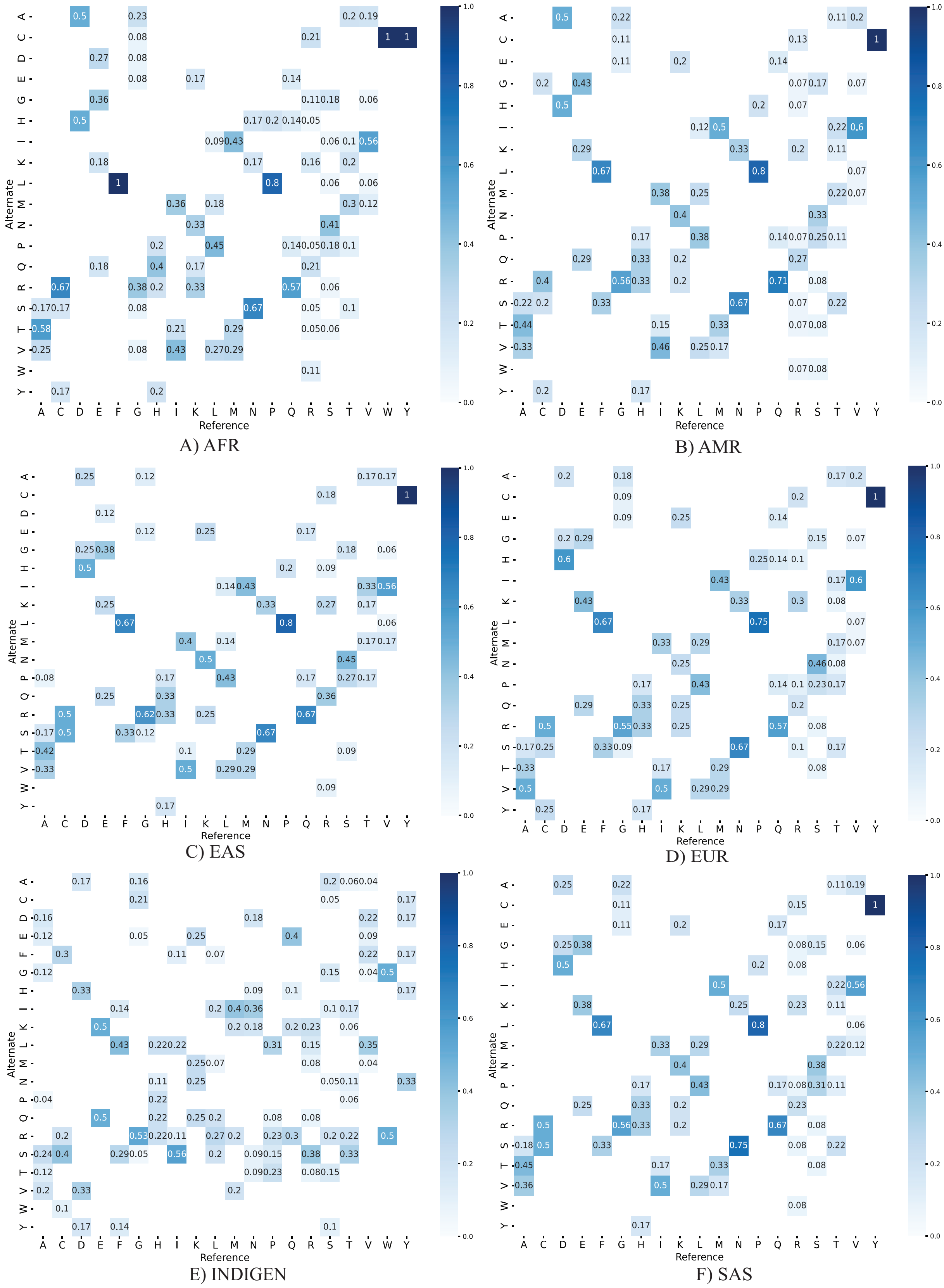
3.1.1 Analyzing trend of amino acid exchanges among diverse ethnicity

To identify the most common variants that contribute to the overall disease burden, we applied an allele frequency filter of 10% to the population data [19]. This resulted in a final set of 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) that were used in subsequent analysis.

Afterwards, we conducted a thorough analysis of the frequency of amino acid substitutions within variant proteins associated with lipid homeostasis and reverse cholesterol transport pathways across six distinct populations, namely African, Admixed American, European, East Asian, South Asian, and Indian. Our primary objective was to investigate prevalent mutational patterns in each population and determine potential differences or similarities in amino acid substitution frequencies for all recorded non-synonymous single nucleotide polymorphisms (nsSNPs) present in the 1,000 Genomes Project (1kG) & IndiGen populations.

We specifically focused on nsSNPs because these variants result in an amino acid change, making them more likely to have a functional impact on the protein's structure and function [17].

To visualize our outcomes, we represented them in a matrix 3.1 & 3.2, with the X-axis represents the amino acid at the reference site, the Y-axis represents its alternate amino acid observed in all six populations, and the value in each cell corresponds



A) AFR

B) AMR

C) EAS

D) EUR

E) INDIGEN

F) SAS

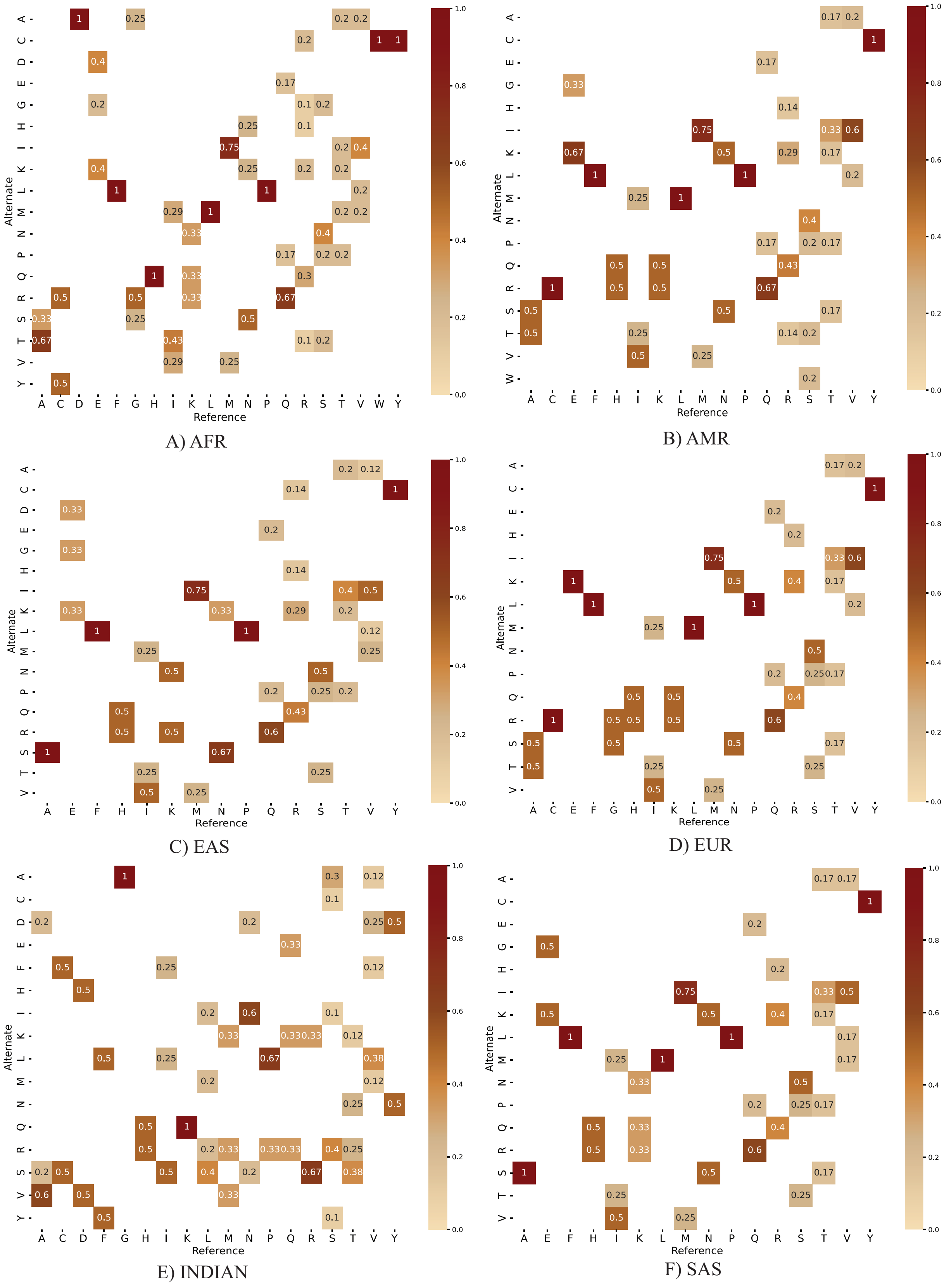
Figure 3.1: Amino Acid Exchange Trend Among Diverse Population

to the frequency of a specific amino acid exchange. The intensity of the cell ranged from darkest to lightest, representing the most frequent to the least frequent exchange, respectively.

Our comparative analysis of all possible amino acid exchanges in the human genome in both the pathways revealed the distinct patterns in the amino acid exchange profiles for different ethnicities. Due to the dissimilarity in the frequency of amino acid exchanges among diverse populations and their mutability differences, the amino acid exchange matrix did not turn out to be balanced and several exchanges were completely absent, however certain others were over-represented. We also observed that the most recurring exchange in one population was rare in another, indicating that the evolution of mutations may differ among ethnic groups or subgroups due to differences in ancestry [19].

In particular, we observed that the SNPs of genes involved in lipid homeostasis pathway for the African population showed a 100% frequency of Phenylalanine (F) being exchanged with Leucine (L) as the most prevalent exchange across all populations, Tryptophan (W) with Cysteine (C) as the population-specific exchange which was contributed by (MECR & LIPC) and ABCA12 genes. The Tyrosine (Y) to Cysteine (C) exchange was the most frequent in all other populations with 100% frequency contributed by ABCG8 gene, except in the Indian population where the exchange frequency of Phenylalanine (F) to Leucine (L) and Tyrosine (Y) to Cysteine (C) ranged from 17% to 43%. In the Indian population Iso-leucine (I) to Serine (S) was found to be most frequent with 56% occurrence frequency which was contributed by NPC1, PNPLA3, TTC39B, APOB and GPAM genes.

In addition to the most frequent substitutions, we also identified several less frequent exchanges that were present in all populations and held significant importance, despite occurred with differing frequency. For instance, we observed a substitution of Proline (P) with Leucine (L) occurred at a frequency range from 30% to 80%, which was contributed by the RBP1, FGFR4, TGFB1, and APOB genes. Furthermore, a Cysteine (C) to Arginine (R) substitution was detected at a frequency range of 20% to 67%, which was contributed by the PEX2, PLD4, APOE, and THADA genes. Although these exchanges are not as prevalent as the most frequent ones, they have the potential to affect protein stability and function.



A) AFR

B) AMR

C) EAS

D) EUR

E) INDIAN

F) SAS

Figure 3.2: Amino Acid Exchange Trend Among Diverse Population in RCT

Thus, in cases where the amino acid resulting from a missense mutation is replaced by a different amino acid with distinct properties, this type of substitution is classified as non-conservative [16], and it has the potential to cause variations in the genetic composition and physical traits between the two populations.

Similarly, when we specifically analyzed the SNPs of genes involved in the Reverse Cholesterol Transport (RCT) pathway we identified that the substitutions of Phenylalanine (F) with Leucine (L), Proline (P) with Leucine (L), and Tyrosine (Y) with Cysteine (C) were consistently frequent in all populations with 100% occurrence frequency except for the Indian population where they ranged from (50% to 67%) and no occurrence for Tyrosine (Y) to Cysteine (C) exchange and Cysteine (C) to Arginine (R) exchange which were found to be prevalent in EUR & AMR population with 50% occurrence frequency in AFR population and no existence in the remaining three population.

The less prevalent substitutions observed during amino acid substitution analysis for lipid homeostasis gene variants found at a higher frequency in amino acid substitutions analysis related to Reverse Cholesterol Transport. This observation suggests that these amino acid substitution are more functionally important in the reverse cholesterol pathway compared to the lipid homeostasis pathway. Hence, supports the importance of specifically studying the reverse cholesterol pathway.

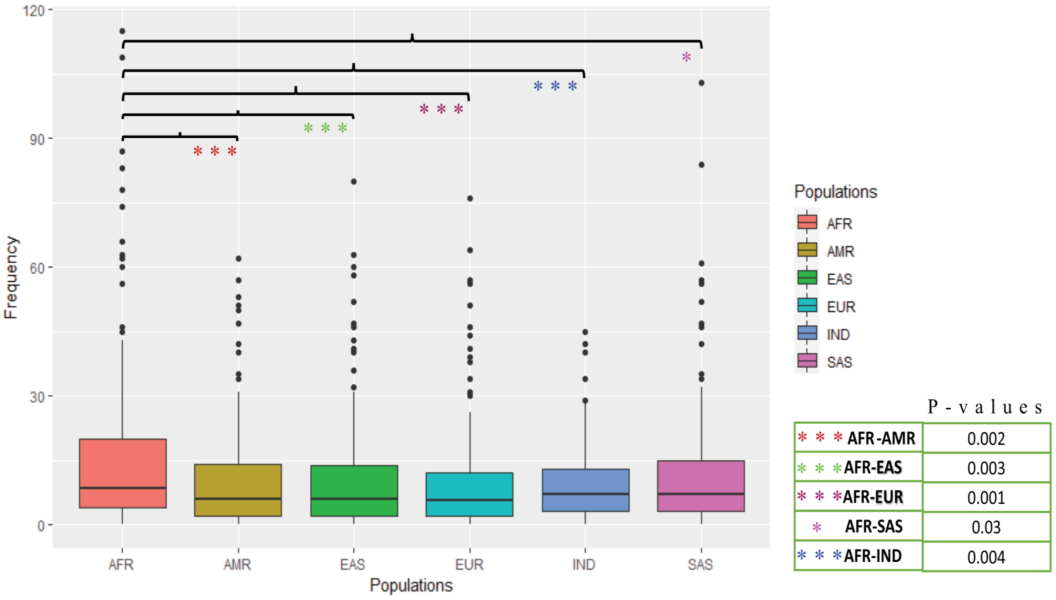
Therefore, again certain substitutions were found to be specific to a particular population. For example, the substitution of Aspartate (D) to Alanine (A) was exclusively detected in the African population, with no reported occurrence in other populations.

The results are indicating that the African population might had undergone more extensive genetic divergence compared to the other populations, leading to the accumulation of distinct amino acid substitutions that are unique to that population. Additionally, it may also suggest that the African population has been subjected to unique environmental pressures or selective forces, leading to the emergence of specific genetic variants that are beneficial for survival in that environment [50].

Moreover, the findings underscore the significance of examining both pathways. The analysis of the Lipid Homeostasis pathway showed that certain amino acid substitutions were disproportionately represented in different populations, while others were entirely absent. Conversely, the examination of the Reverse Cholesterol Transport pathway indicated that the substitutions of Phenylalanine (F) with Leucine (L), Proline (P) with Leucine (L), and Tyrosine (Y) with Cysteine (C) were consistently frequent in all populations except for the Indian population. Furthermore, less frequent substitutions observed during amino acid substitution for lipid homeostasis gene variants were also observed more frequently in amino acid substitutions related to Reverse Cholesterol Transport. This highlights the presence of both pathway-specific and population-specific substitutions, implying that diverse populations exhibit distinct genetic characteristics.

3.1.2 Statistical Analysis

To further examine the potential genetic differences between populations in terms of amino acid substitution frequencies, a statistical comparison was performed. Since amino acid substitution involves comparing frequencies across multiple populations, the statistical tests were used to determine whether observed differences in substitution frequencies were statistically significant or due to chance.



	P-values
* * * AFR-AMR	0.002
* * *AFR-EAS	0.003
* * *AFR-EUR	0.001
* AFR-SAS	0.03
* * *AFR-IND	0.004

Figure 3.3: Statistical Analysis

This involved representing the data using box plots 3.3, with the occurrence frequency of amino acid substitutions plotted against each population on the x-axis and the y-axis, respectively. To analyze this count variable data & to compare the frequency of amino acid substitutions between pairs of populations, a non-parametric

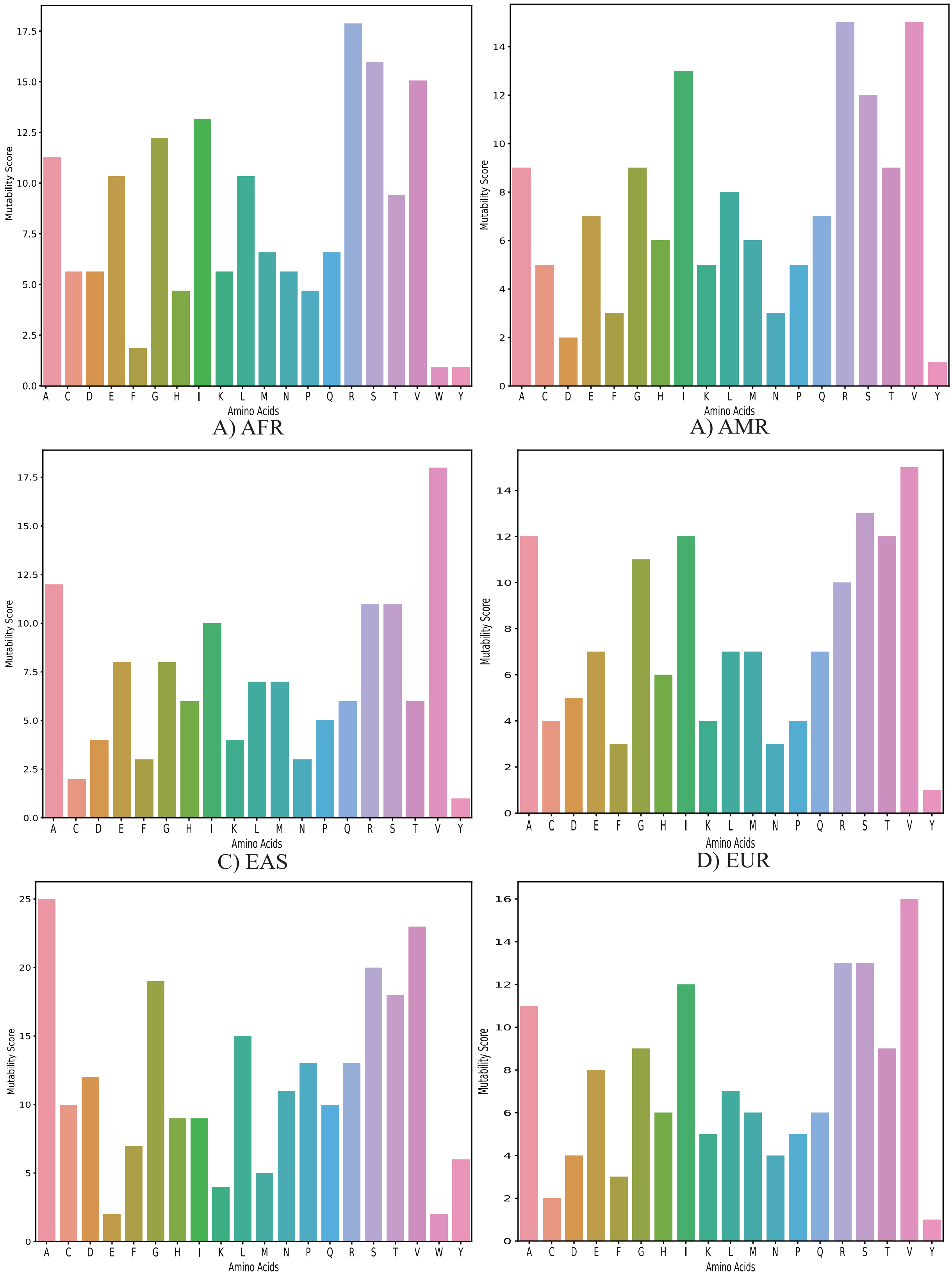
Mann-Whitney U-test was used. A p-value of less than 0.05 was set as the level of significance for determining whether a statistically significant difference existed. Thus, the results revealed that the AFR population had a distinct pattern of amino acid substitution frequency compared to the other populations, indicating unique evolutionary history. Hence, these results support the observation of amino acid substitution analysis, which showed that certain amino acid substitutions were unique to the AFR population and were not present in the other populations.

3.1.3 Mutability

With the purpose to determine which amino acids are most prone to mutations in these metabolic pathways we investigate the mutability of different amino acids across populations from diverse geographic regions. The bar plots was generated to illustrate the mutational rates 3.4 & 3.5, with the amino acids plotted on the x-axis and the corresponding mutability scores displayed on the y-axis. The results highlight that different populations exhibit different levels of mutability for the 20 amino acids in both pathways.

In the African population, Arginine (R) was found to be the most mutable amino acid in both pathways. Conversely, in the American population, Valine (V) and Arginine (R) exhibited the highest mutation rate in the lipid pathway, while only Arginine (R) was highly mutable in the reverse cholesterol pathway. In the East Asian population, Valine (V) was the most mutable amino acid in both pathways. For the European population, Valine (V) had the highest mutation rate in the lipid pathway, whereas Threonine (T) was the most mutable in the reverse cholesterol transport pathway. In the Indian population, Alanine (A) was the most mutable in the lipid pathway, while Serine (S) had the highest mutation rate in the reverse cholesterol transport pathway. In the South Asian population, both Threonine (T) and Valine (V) had the highest mutation rates in the reverse cholesterol pathway, while only Valine (V) was the most mutable amino acid in the lipid pathway.

Furthermore, codon usage bias was observed in certain populations, with lower mutation frequencies occurring for amino acids with a greater number of codons compared to those encoded by fewer codons [51]. Overall, these results emphasize that different



A) AFR

A) AMR

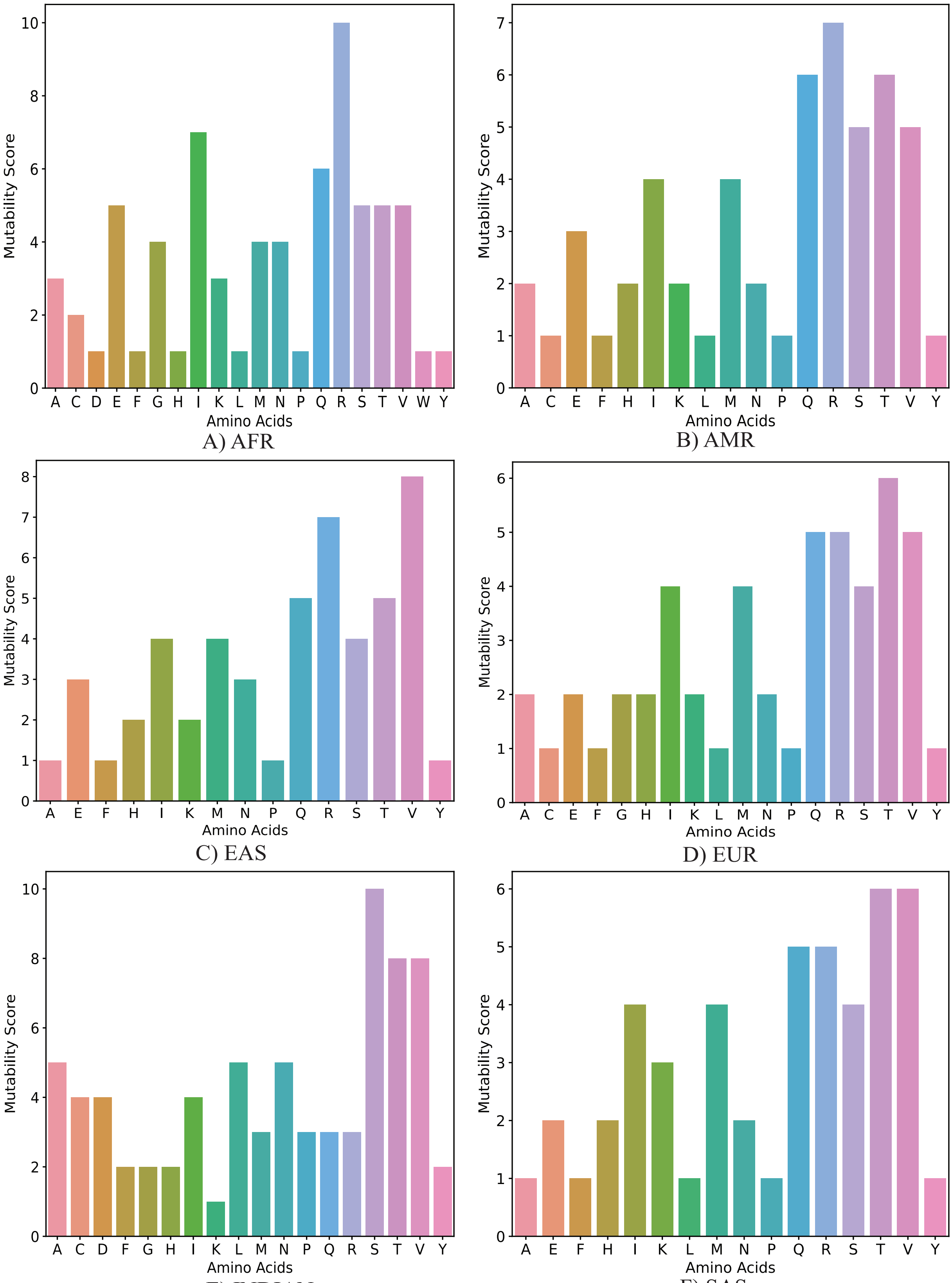
C) EAS

D) EUR

E) INDIAN

F) SAS

Figure 3.4: Analyzing mutability rate across different populations for Lipid Homeostasis variants



A) AFR

B) AMR

C) EAS

D) EUR

E) INDIAN

F) SAS

Figure 3.5: Analyzing mutability rate across different populations for RCT variants

populations exhibit distinct patterns of amino acid mutability.

3.1.4 Chemical Shift Analysis

Consequently, these amino acid substitutions can induce alterations in the chemical characteristics of the proteins, which could impact their structure and function. To elucidate the nature whether these amino acid substitutions involve inter-class or intra-class conversions, a chemical shift analysis was conducted. The amino acids were categorized into eight distinct chemical classes based on their functional groups and chemical properties, namely Aliphatic (G, A, L, V, I), Hydroxyl (S, T), Cyclic (P), Aromatic (F, Y, W), Basic (K, R, H), Acidic (D, E), Sulpho (C, M), and Amides (N, Q). To examine the inter-class conversion and intra-class conversion of these amino acid classes, we developed a matrix 3.6 & 3.7 where the reference amino acid classes were arranged in columns and the alternate amino acid classes were arranged in rows. The matrix entries indicate the frequency of conversions occurring between the reference and alternate amino acid classes.

Based on our analysis of chemical shifts we had observed that inter-class conversions occur more frequently than intra-class conversions in all six populations among both pathways studied. This was evidenced by a higher number of non-zero values off the diagonal compared to on the diagonal.

As our study primarily focuses on the most frequently occurring amino acid substitutions, so we had considered the most frequently observed mutation identified from our amino acid substitution analysis (FL, YC, CR, and PL) in our subsequent analysis. Notably, we observed that same substitutions were present in the reverse cholesterol pathway, suggesting their possible importance in both pathways. Hence, these substitutions were considered to assess the effects of these variants on both the lipid and reverse cholesterol pathways.

We investigated their properties by analyzing the chemical shift analysis results. Our observations revealed that all of these substitutions involved in inter-class conversions except F to L substitution involved in intra-class conversion, indicating their potential to cause significant changes in protein function [52].

In terms of polarity changes, we observed that the substitutions from Phenylalanine

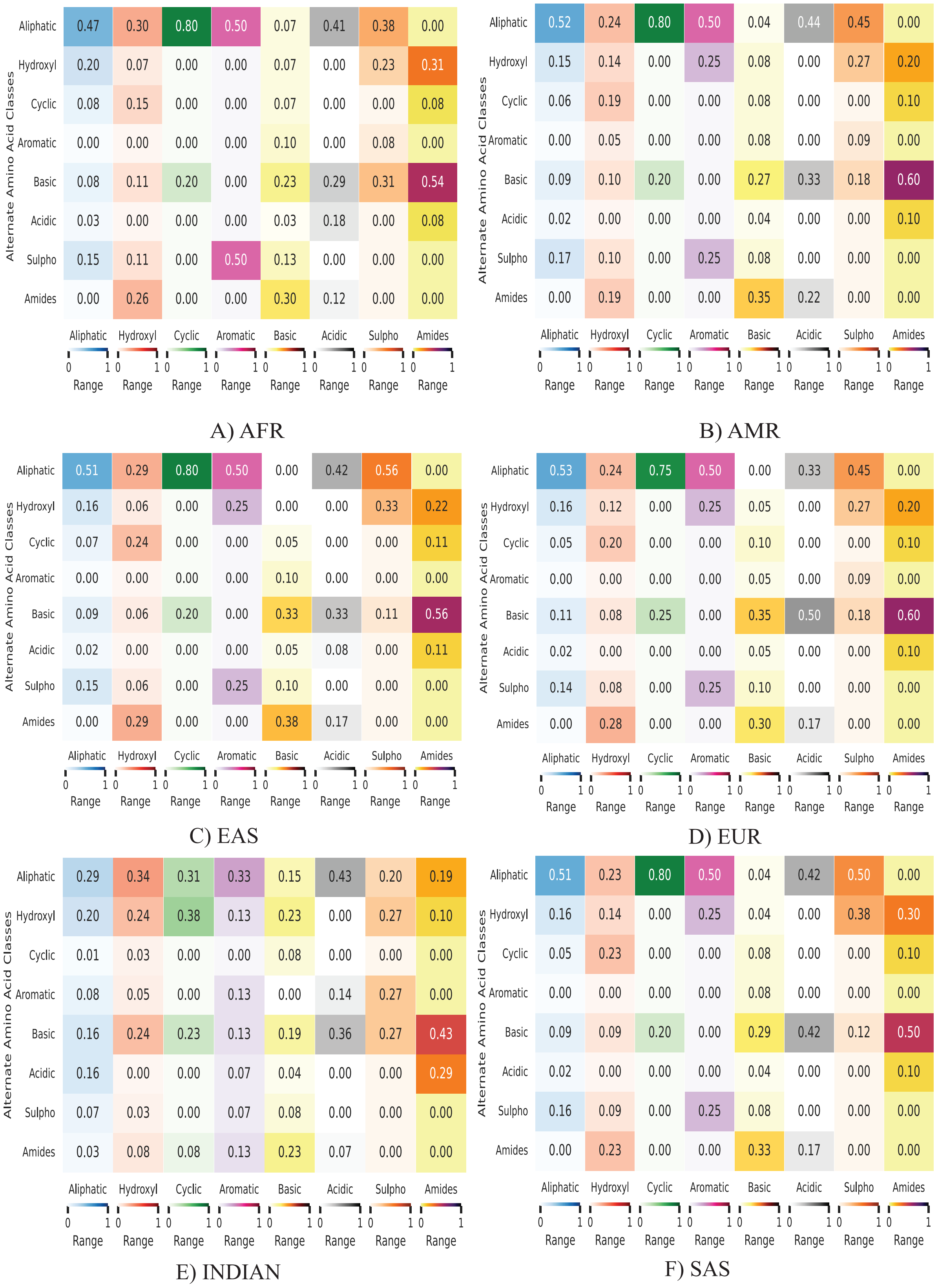
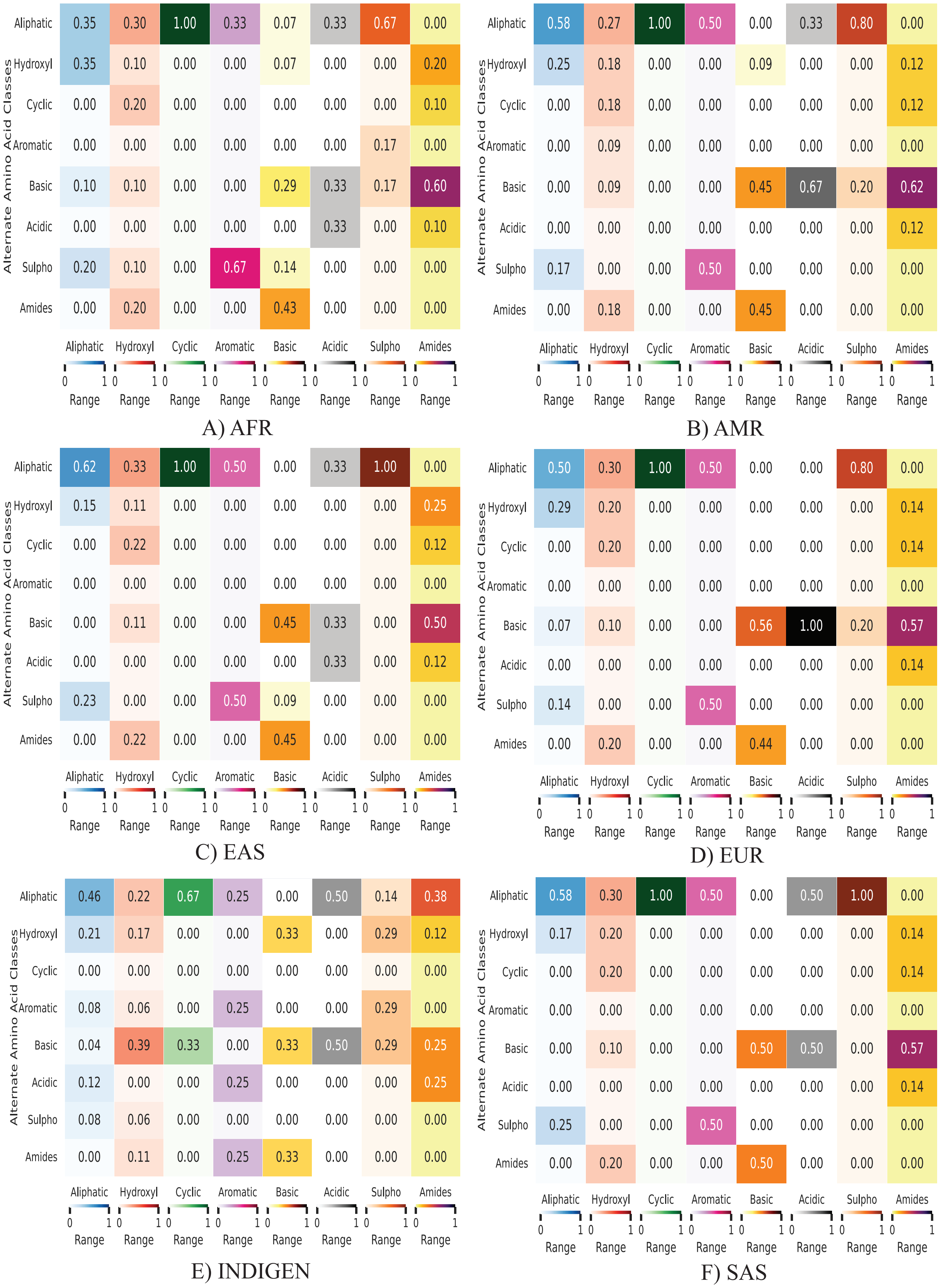


Figure 3.6: Chemical Shift Analysis



A) AFR

B) AMR

C) EAS

D) EUR

E) INDIGEN

F) SAS

Figure 3.7: Chemical Shift Analysis for RCT

(F) to Leucine (L) and Tyrosine (Y) to Cysteine (C) entailed the replacement of relatively polar amino acids with less polar or non-polar amino acids, resulting in a decrease in polarity. According to reports, a decrease in polarity resulting from an amino acid substitution can have implications as seen in sickle cell anemia [53] or in the gene that encodes the tumor suppressor protein p53 [54].

Conversely, the Cysteine (C) to Arginine (R) substitution involved the replacement of a relatively non-polar amino acid with a polar amino acid, indicating an increase in polarity. According to several studies, amino acid substitutions can have significant implications for the APOE gene, which encodes a protein involved in lipid metabolism and transport. One particular polymorphism, referred to as epsilon-4, results in the substitution of a non-polar amino acid, Cysteine (C), with a polar amino acid, Arginine (R), at position 112 of the protein. This substitution increases the polarity of the protein surface, which could impact interactions with other molecules and contribute to the onset of Alzheimer's disease [55]. Similarly, in cystic fibrosis, a mutation leads to the deletion of a single amino acid, Phenylalanine (F), at position 508 in the protein. This deletion enhances the polarity of the protein surface, which may result in increased interactions with water molecules, reduced stability, and decreased efficiency of protein folding, ultimately contributing to the development of cystic fibrosis [56]. Finally, the Proline (P) to Leucine (L) substitution involved the replacement of two relatively non-polar amino acids, resulting in no significant change in polarity. While this substitution may not alter the polarity but can still have implications on the protein, depending on the specific amino acids involved and the location of the substitution within the protein.

Conclusively, our results suggest that amino acid substitutions were more common between different classes of amino acids than within the same class, indicating that substitutions were more likely between amino acids of different structural and chemical classes and overall the substitutions we identified result in diverse changes in amino acid polarity.

3.1.5 Tracing the variants impact on Protein Domain

Initially, we investigated the presence of variants in/out of a conserved protein domain since the variants within the domain are more probable to impact the structure, stability, and function of the protein, the protein domain analysis was executed. The variants present in the genomes of Indigen and 1kG projects from different individuals for European, American, African, East Asian, South Asian and Indian populations were classified into pre-domain, post-domain, and within the domain regions based on their position. A polar bar plot was generated to represent the distribution of variants in each population across these regions 3.10. The angle of each bar corresponds to the proportion of variants in the pre-domain, post-domain, and within the domain regions. The plot displays the variant counts for the African (grey), American (yellow), East Asian (cyan), European (violet), South Asian (pink), and Indigen (light green) populations.

The analysis of SNVs across all populations showed that there were fewer SNVs in pre-domain regions. This indicates that there was a bias for SNVs to occur within protein domains or post-domain regions in both the 1000 Genome and IndiGen data. The distribution of SNVs in post-domain and within-domain regions was similar across all populations. In the Indian population, the majority of variants (91) fell within the protein domain, while 136 fell within the post-domain region, with only two variants observed in the pre-domain region. The African population had the highest number of SNPs (155) in the post-domain region. Additionally, all of the prevalent exchanges we identified were located within the domain region for these genes.

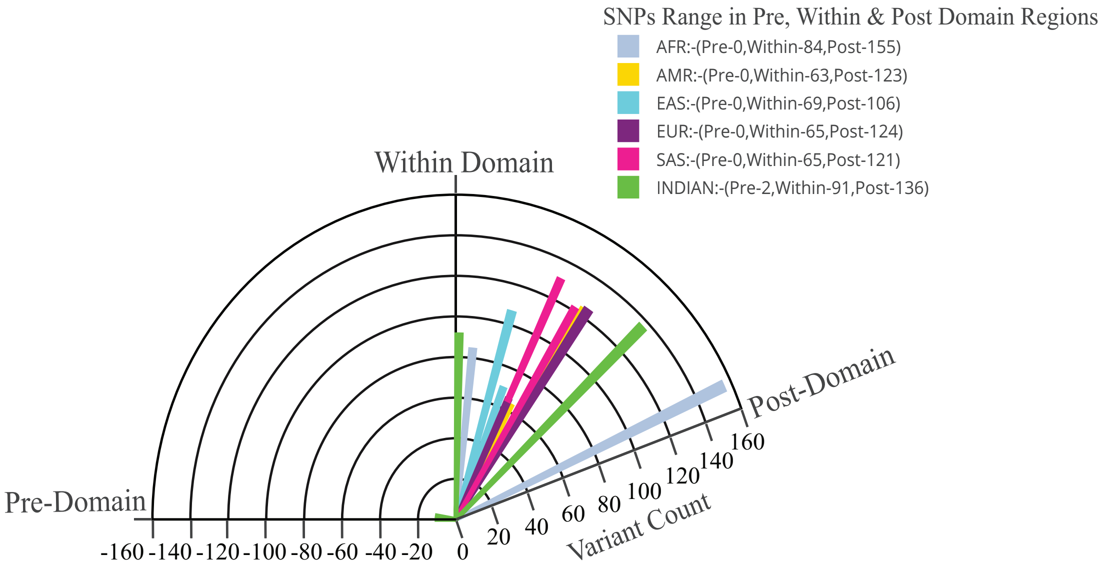


Figure 3.8: Protein Domain Analysis

3.2 Structural Analysis

After comprehending the impacts of variations at the sequence level, it becomes essential to gain a more profound understanding of how genetic variants influence the protein's stability, structure, and function. The presence of a single nucleotide variant (SNV) in functionally important genes could be pathogenic since it might alter the secondary, tertiary, or quaternary structure of a protein, impacting its interactions with other molecules and its capacity to perform its normal functions. Conducting structural analysis investigations can offer insights into how these non-synonymous variants influence protein stability, solvent accessibility, and hydrophobicity, as well as how they affect protein structure and function in both conserved and flexible regions. Only human proteins for which a complete PDB structure was known were included in the 3-D study to enable accurate analysis of 3-D characteristics.

To investigate the impact of the identified variants, we applied several filters to the master data used for sequence-level analysis, including a minimum allele frequency of 10%, availability of at least one protein crystal structure, sequence coverage of at least 70%, and SNV coverage to the crystal structure. After applying these filters, we obtained a subset of data consisting of 31 variants in 20 genes for the African population, 22 variants in 15 genes for the American population, 24 variants in 17 genes for the East Asian population, 22 variants in 16 genes for the European population, 53 variants in 28 genes for the Indian population, and 22 variants in 15 genes for the South Asian population. This filtered data was then used to analyze the structural implications of these non-synonymous SNPs.

3.2.1 Integrating nsSNP data with 3D protein structures: Mapping and Analysis

Residue type variations at specific protein sites are influenced by individual genetic diversity [57]. To understand how non-synonymous SNPs affect protein structure, each SNP was mapped to its corresponding amino acid in the protein structure. A single amino acid change can impact the entire peptide chain, resulting in a new protein variant. This mutation affects every step of the cellular reproduction process, as the original protein is necessary for reproduction to occur. We generated homology models of the variants using Modeller [Version-10.3] and mapped the non-synonymous SNPs using the original protein structure as a template.

3.2.2 Protein Secondary Structure & Solvent Accessible Area

To analyze the secondary structure of the protein variants we used the Dictionary of Secondary Structure of Proteins (DSSP). Across all populations, a higher percentage of SNVs mapped to regions other than alpha helices and beta sheets 3.11. Few studies have shown that the tolerance for mutations varies significantly in alpha helices, beta sheets and other regions of the protein [22].

Our analysis also revealed that certain variant models exhibited changes in the secondary structure of their amino acid residues upon accumulating mutations. Specifically, we observed that residue substitutions in CBR4 and THOC5 variants (L70M and V525I) resulted in changes in secondary structure, where they were found in Alpha-Helix and Others (include bends, bulges, loops, and non-repetitive structures (turns, coils, loops, and irregular helices)) in the mutants but in Turn and Strand in the wild-type 3.9. However, for other gene variants, the secondary structure remained unchanged despite accumulating mutations.

In the Indian population, we observed changes in the secondary structure of specific gene variants due to accumulated mutations. These include the ABCA1 variant L681I, ABCB11 variants at positions 1029 and 591 (T1029A and N591K), ABCC8 variant V563F, CBR4 variant resulting in the substitution of Q for L at position 70, LPL variant T388N, MVK variant D170V, NPC1 variants K643Q and V130L, PCSK9 gene variant V460G, and POLD1 variant where Iso-leucine (I) replaces Threonine (T) at position 495.

Besides this, we were focused on the most frequent amino acid exchanges, which included FL, YC, CR, and PL in genes (such as MECR, ABCG8 etc) from the Lipid Homeostasis and Reverse Cholesterol Transport pathways, were examined. It must be noted that the ABCG8 and APOE genes have well-established roles in the RCT pathway, and mutations in these genes can have significant implications for cholesterol metabolism.

The substitution of Phenylalanine (F) to Leucine (L) was contributed by both the

MECR and LIPC genes, while the Tyrosine (Y) to Cysteine (C) exchange was contributed by the ABCG8 gene. However, since the crystal structure of the LIPC gene was unavailable in RCSB-PDB, we only analyzed the impact of the Phenylalanine (F) to Leucine (L) substitution at position 96 over the MECR gene. Despite the high allele frequency (0.7231, 0.7334, 0.6943, 0.8797, 0.5625) in African, American, South Asian, European, and East Asian populations, our results showed that this substitution had no effect on the protein's secondary structure. In the Indian population, this exchange was observed, but we could not analyze its impact on the crystal structure since the contributing genes (TSPO, MALRD1, and CHPT1) had no available structure in RCSB-PDB.

On the other hand, a Y to C substitution at position 54 of the ABCG8 protein was identified in all populations except for the Indian population. Despite the accumulation of mutation, this substitution did not appear to cause any changes in the protein's secondary structure.

The substitution of C with R was attributed to the PEX2, PLD4, APOE, and THADA genes. Unfortunately, we could not analyze the implications of this substitution for the PEX2 and PLD4 genes as their crystal structures were not available in the RCSB-PDB database. For THADA, although a crystal structure with PdbId-5T6Y was available, its sequence coverage was below the threshold of 70%, which was one of our filter conditions. Therefore, we did not consider the crystal structure of the THADA gene for our analysis. Consequently, we investigated the impact of the C to R substitution on the APOE gene, which was observed at position 130. However, our analysis revealed no significant changes in the protein's secondary structure. Therefore, this substitution was not observed in the Indian population.

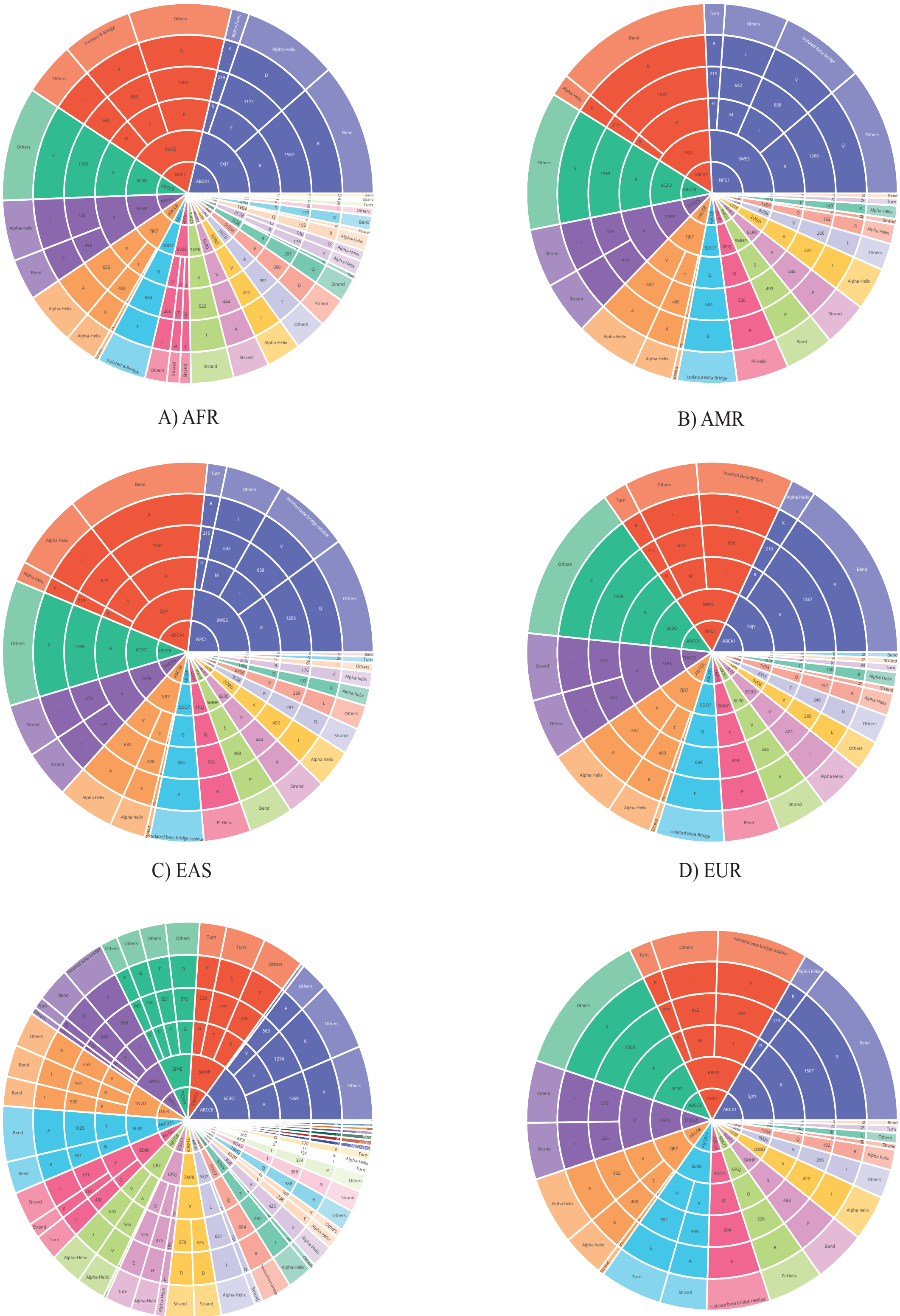
The P to L substitution was attributed to RBP1, FGFR4, TGFB1, and APOB genes. Unfortunately, the crystal structure of APOB was not available in RCSB-PDB, and the FGFR4 gene's crystal structure had sequence coverage less than 70%. Moreover, the reference amino acid information for the RBP1 gene variant did not match with the UniProt native sequence, and the substitution in the TGFB1 gene was located in the signal peptide region, making it unsuitable for homology modeling.

Overall, we found that specific amino acid substitutions in certain gene variants can cause changes in the protein's secondary structure, as evidenced by our observations of the CBR4 and THOC5 variants across different populations. However, despite accumulating mutations in other gene variants, we did not observe any changes in the secondary structure. Our analysis was focused on the most frequent amino acid substitutions in genes from the Lipid Homeostasis and Reverse Cholesterol Transport pathways, such as MECR, ABCG8, and APOE. We determined that the substitution of Phenylalanine (F) with Leucine (L) at position 96 of the MECR gene and the Tyrosine (Y) to Cysteine (C) substitution at position 54 of the ABCG8 protein did not have any significant impact on the protein's secondary structure. Unfortunately, we could not analyze the implications of some substitutions due to unavailability of crystal structures or insufficient sequence coverage.

Following the secondary structure analysis, it becomes essential to evaluate the solvent accessibility of the identified nsSNPs. For this purpose, we employed the Naccess tool to perform a solvent accessibility analysis on the side chain. The analysis calculated the Relative Solvent Accessibility of Total-Side (RSA) for each nsSNP 3.10, allowing us to categorize them as either Buried or Exposed, with an additional option for Either Buried or Exposed state, based on a threshold of [<=30% for Buried, >=70% for Exposed, and >30% & <70% for Either Buried or Exposed]. RSA represents the percentage of protein residues exposed in the native structure. Higher RSA percentages indicate greater surface exposure, while lower percentages indicate more buried residues. BorE describes whether the mutation or SNP is situated inside the protein or on the protein's surface.

The study revealed that in the AFR population, the Buried class of SNVs had a distribution of 31% on the Alpha Helix, 22% on the Beta Sheets, and 37% on other regions of the protein. Conversely, for the Exposed class of SNVs, 50% were present on the Alpha Helix and other regions (turn, coils, bend, loops), while the proportion was lower on the Beta Sheets. Additionally, the probability of a residue being either an exposed or buried residue on the Alpha Helix was 18%, on the Beta Sheets it was 46%, and on other regions it was 36% 3.11.

Similar patterns were observed in the Buried class of SNVs in the AMR population, where 34% were on the Alpha Helix, 33% on the Beta Sheets, and 33% on other regions. The remaining 34% of SNVs on the Alpha Helix, Beta Sheets, and other regions were likely either exposed or buried residues. The SNVs in the Exposed class were



A) AFR

B) AMR

C) EAS

D) EUR

E) INDIAN

F) SAS

Figure 3.9: Sunburst Plot representing the availability of the variant in the secondary structure

predominantly found in other regions.

The Buried class of amino acids in the EAS population showed that 31% were situated on the Alpha Helix, 38% on the Beta Sheets, and 31% on other protein regions. Conversely, SNVs in the Exposed group were most frequently located in other regions with 40% on the Alpha Helix, 20% on the Beta Sheets, and the remaining 40% potentially being Exposed or Buried variant residues.

In the EUR population, 27% of the Buried class of amino acids were found on the Alpha Helix, 40% on the Beta Sheets, and 33% on other regions of the protein. All SNVs in the Exposed category were most likely located in other regions, with 17% on the Alpha Helix, 33% on the Beta Sheets, and the remaining 50% potentially harboring an Exposed or Buried variant residue.

For the Buried class of SNVs in the Indian population, 24% were positioned on the Alpha Helix, 33% on the Beta Sheets, and 43% on other protein regions. In the Exposed group, 11% were mapped to the Alpha Helix, 6% to the Beta Sheets, and 83% to other regions. The likelihood of finding a residue on the protein surface as an exposed residue or buried inside the protein's core was 36% for the Alpha Helix region, 14% for the Beta Sheets, and 50% for other regions.

Finally, in the SAS population, 25% of the Buried class of amino acids were found on the Alpha Helix, 50% on the Beta Sheets, and 25% on other regions of the protein. SNVs in the Exposed category were most frequently located in other regions, and 38% of SNVs were on the Alpha Helix, 12% on the Beta Sheets, and the remaining portions had a 50% likelihood of harboring an Exposed or Buried variant residue.

Therefore, the Buried class of nsSNVs exhibits a significant range of 22% to 50% in the Beta-Strand secondary structure. When located in the beta-strand, which serves as a key structural element of proteins, these nsSNVs may disrupt protein stability and structure, potentially leading to functional loss or modification. On the other hand, the Exposed class indicates that a high proportion of nsSNVs, ranging from 50% to 100%, are found in other regions. The largest number of variations in this class, with a proportion of occurrence between 33% to 50%, are detected in either the buried or exposed class, suggesting that these regions are more susceptible to mutations that affect protein function.

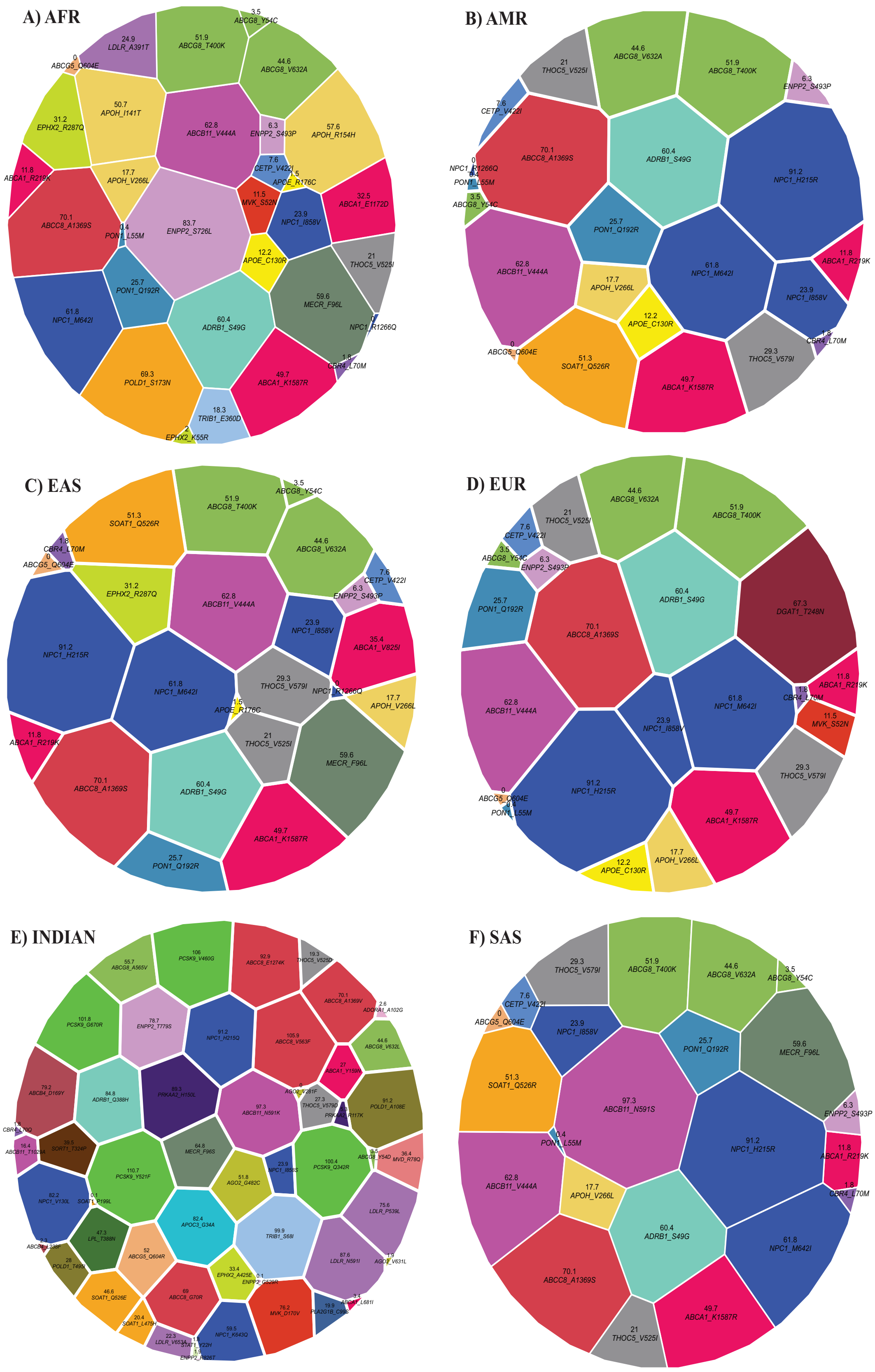
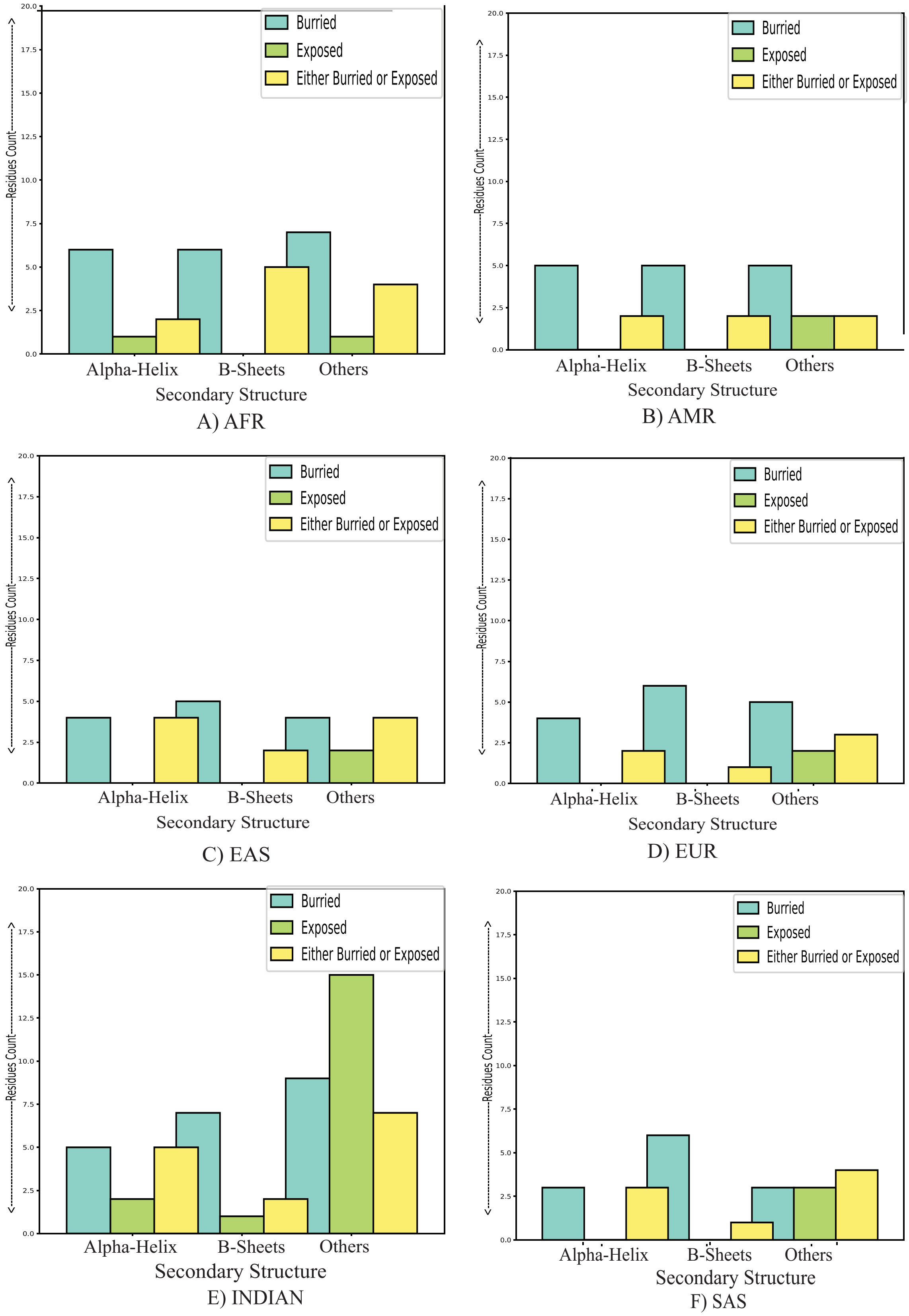


Figure 3.10: Voronoi Treemap representing the Relative Solvent Accessible Area by Variants



A) AFR

B) AMR

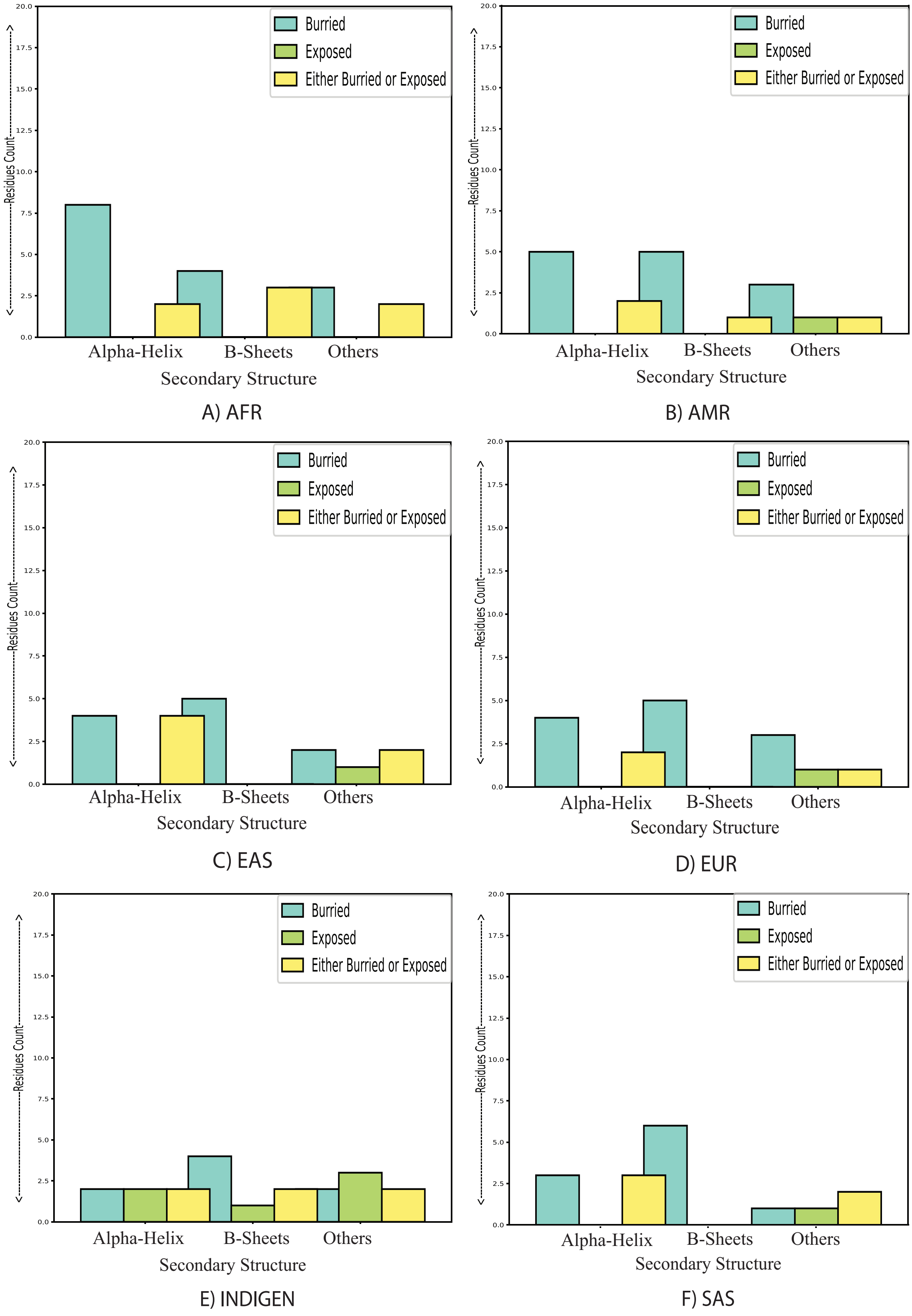
C) EAS

D) EUR

E) INDIAN

F) SAS

Figure 3.11: DSSP-Naccess Analysis



A) AFR

B) AMR

C) EAS

D) EUR

E) INDIGEN

F) SAS

Figure 3.12: DSSP-Naccess Analysis in RCT

In the context of the Reverse Cholesterol Pathway 3.12, analysis of nsSNVs reveals that 53% of SNVs classified as Buried in the AFR population are located on the Alpha Helix, 27% on the Beta Sheets, and 20% in other regions, with none belonging to the Exposed class. The probability of finding an nsSNV in the Alpha Helix region, either as an exposed residue on the protein surface or as a buried residue in the protein core, is 28%, while it is 43% for the Beta Sheets and 28% for other regions.

In the AMR population, the buried class of SNVs consisted of 39% of SNVs in the Alpha Helix, 38% in the Beta Sheets, and 23% in other regions. Within the alpha helix region, 50% of SNVs were found to be buried, and 50% exposed, while the same was true for other regions. Notably, the Exposed SNVs were mainly located in regions other than the Buried/Exposed class.

Similarly, the buried class of amino acids in the EAS population showed 36% of residues in the alpha-helix, 46% in the beta-sheets, and 18% in other regions. The Exposed SNVs were mainly found in regions other than the Buried/Exposed class, with 67% in the alpha-helix region and the remaining 33% in regions that could harbor either Exposed or Buried variant residues.

In the EUR population, 33% of residues in the buried class were mapped to the alpha-helix, 42% to the beta-sheets, and 25% to other regions. All SNVs in the Exposed category were located in regions other than the Buried/Exposed class, with 67% in the alpha-helix region and the remaining 33% in regions that could contain either Exposed or Buried variant residues.

For the Indian population's Buried category, 25% of SNVs mapped to the Alpha Helix, 50% to the Beta Sheets, and 25% to other regions. The Exposed group included 33% of SNVs in the alpha helix, 50% in the beta sheets, and 17% in other regions. The probability of finding a residue exposed or buried was similar across all three regions: 34% for the alpha helix, 33% for the beta-strand, and 33% for other regions.

Finally, in the SAS population, the buried class of amino acids consisted of 30% of residues in the alpha-helix, 60% in the beta-sheets, and 10% in other regions. The Exposed SNVs were primarily found in regions other than the Buried/Exposed class, with 60% in the alpha-helix region and the remaining 40% potentially containing either Exposed or Buried variant residues.

Encapsulating, the Buried class in RCT displayed a variable range of nsSNVs in the Beta-Strand secondary structure, with a maximum number ranging from 27% to 60%. Conversely, the Exposed class indicated that other regions could have 17% to 100% of nsSNVs. The alpha helix regions had the highest number of variations in either the Buried or Exposed class, ranging from 28% to 67% 3.12.

Furthermore, we observed alterations in solvent accessibility for certain amino acid residues due to the accumulation of mutations. Upon comparing the wild-type and mutant models, specific residue substitutions resulted in increased or decreased accessibility. Notably, in some populations, the V444A and V632A variants of ABCB11 and ABCG8, respectively, showed the largest differences in solvent accessible area between the wild-type and mutant models. Specifically, the ABCB11 variant exhibited decreased accessibility, while the ABCG8 variant showed increased accessibility after accumulating mutations.

However, these findings were not consistent across all populations. In the Indian population, the ABCC8 and ABCG8 variants (V563F and V632L) showed the most significant changes in relative solvent accessibility (RSA) percentage between wild-type and mutant variants. The ABCC8 variant exhibited a notable decrease in accessibility, while the ABCG8 variant showed an increase in accessibility after accumulating mutations. Conversely, in the European population, different observations were reported. For instance, the DGAT1 and ABCG8 variants V632A and T248N demonstrated the largest differences between wild-type and mutant, with the V632A variant leading to increased accessibility, while the T248N variant resulted in decreased accessibility.

It is important to note that certain wild-type residues exhibit a low percentage of accessibility, indicating their buried nature, while others display a high percentage, implying their exposed nature. This indicates that residue accessibility can significantly vary in a protein structure.

In this study, we found that the Y54C and C130R mutations in the ABCG8 and APOE genes, respectively, resulted in an increase in accessibility (by 6.5% and 18.4%) of the mutated residue. This increased accessibility could potentially impact the interaction of these proteins with other molecules involved in the RCT pathway, leading to changes in cholesterol transport and metabolism.

Similarly, the F96L mutation in the MECR gene led to a decreased accessibility (by 12.5%) of the mutated residue. This decreased accessibility could potentially affect the function of MECR and lead to changes in very long chain fatty acid synthesis and subsequent lipid homeostasis.

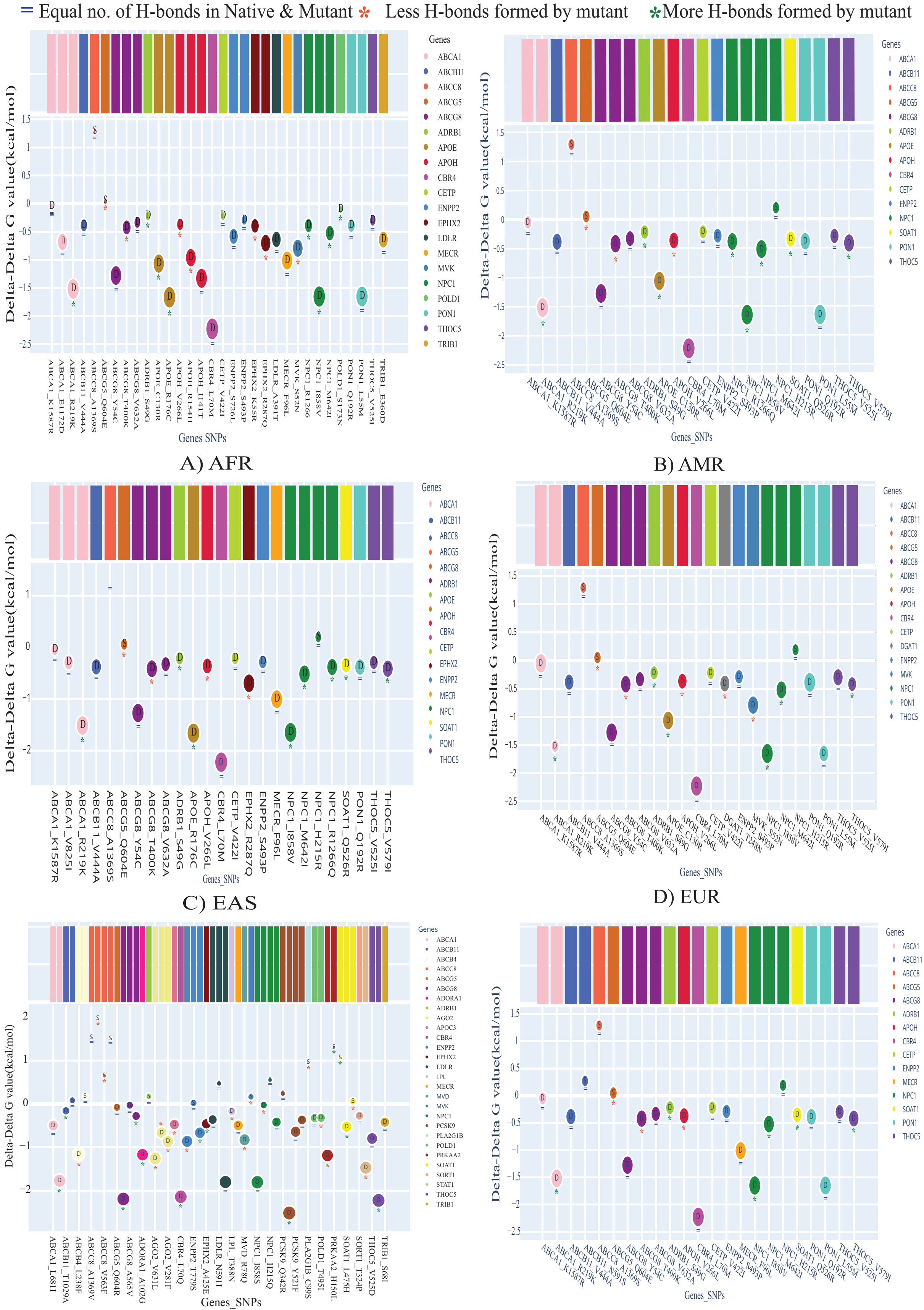
It can be concluded that the Buried class of nsSNPs had a significant range of occurrence in the Beta-Strand secondary structure, which serves as a key structural element of proteins. On the other hand, the Exposed class of nsSNPs was found to be located predominantly in other regions, indicating that these regions are more susceptible to mutations that affect protein function. In the context of the Reverse Cholesterol Pathway, the study revealed that nsSNPs in the Buried class were more likely to be located on the Alpha Helix and Beta Sheets regions, while nsSNPs in the Exposed class were not present in the AFR population.

3.2.3 Assessing the Effects of nsSNPs on Protein Stability and Hydrogen Bond Network

Based on our analysis of the potential consequences of SNPs on solvent accessibility, we aimed to investigate the impact of these mutations on protein stability. To achieve this, we employed the Dynamut2 API to perform stability analysis. By examining the changes in $\Delta\Delta$G values, measured in kcal/mol, we sought to determine whether the SNP-induced impacts would be stabilizing or destabilizing for the gene 3.13.

Our results, which varied across different populations, indicated that 30 out of 31 variants in the AFR population, 20 out of 22 in the AMR population, 20 out of 22 in the EUR population, 22 out of 24 in the EAS population, 37 out of 53 in the Indigen population, and 18 out of 22 in the SAS population had negative $\Delta\Delta$G values, indicating destabilization caused by point mutations 3.15. These findings suggest that such mutations may alter the structure and function of the protein, potentially leading to a range of deleterious effects.

In accordance with our findings, positive $\Delta\Delta$G values were observed for certain variants in different populations. Specifically, the AFR and EUR populations had 2-2 variants, while the AMR and EAS populations had 3-3 variants, the Indigen population had 16 variants, and the SAS population had 4 variants, all indicating increased



A) AFR

B) AMR

C) EAS

D) EUR

E) INDIGEN

F) SAS

Figure 3.13: Mutational effect over Protein stability and H-bond Network

structural flexibility resulting from the mutations.

When examining variants in the Reverse Cholesterol Pathway, we found that negative $\Delta\Delta$G values were prevalent, signifying destabilization of the protein structure. Specifically, the AFR population had 19/20 variants, the AMR and EAS populations had 16/18 variants, the EUR population had 14/16 variants, the Indian population had 16/20 variants, and the SAS population had 14/16 variants with destabilizing effects 3.14.

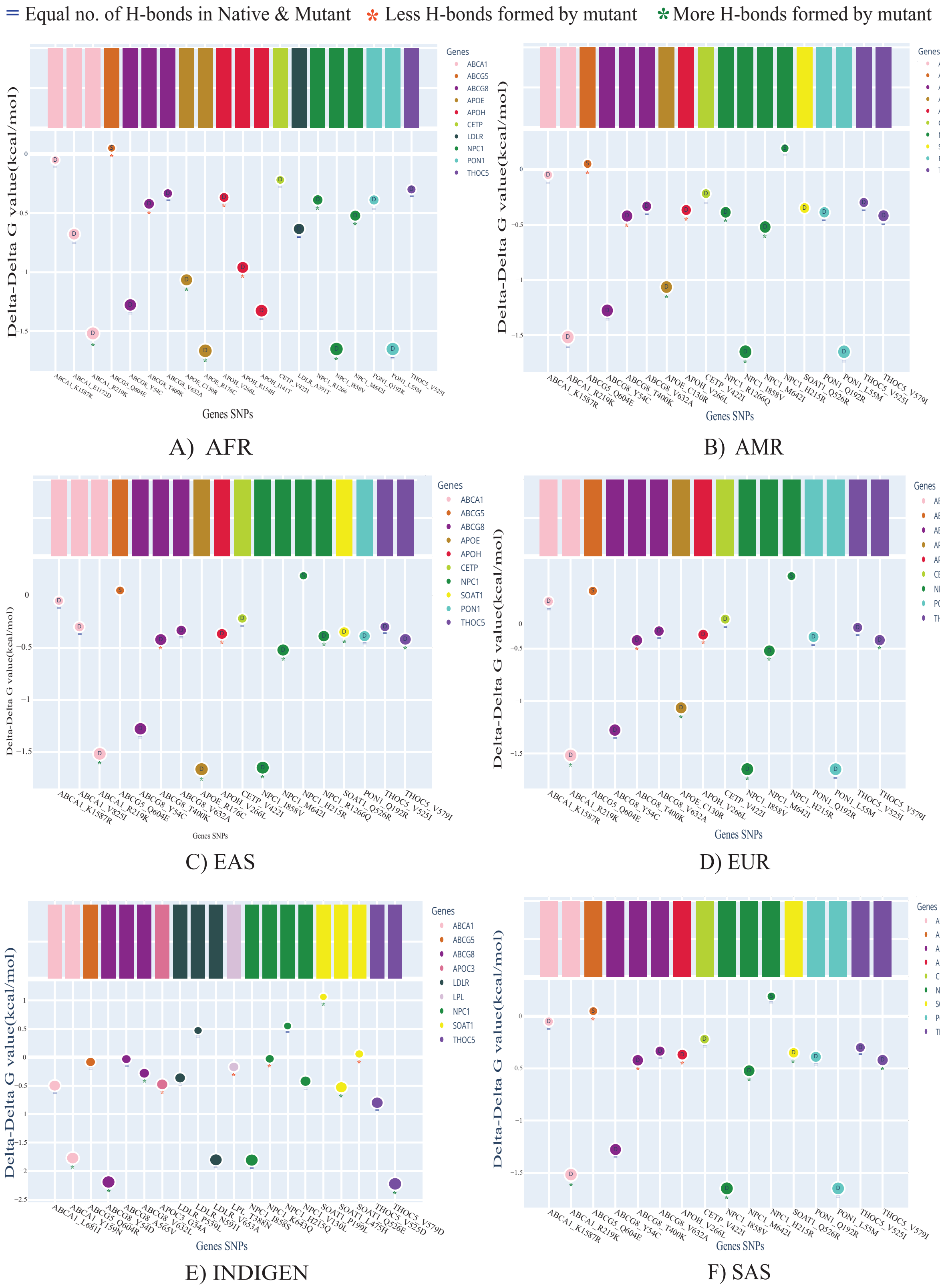
However, a few variants had a stabilizing effect after the incorporation of non-synonymous variants. In particular, the AFR population had 1 variant, while the AMR, EAS, EUR, and SAS populations had 2-2 variants, and the Indian population had 4 variants with positive $\Delta\Delta$G values.

The study demonstrated that the majority of SNPs in genes had a destabilizing effect on the protein structure, with only a few exhibiting a stabilizing effect. The extent of destabilization varied across six populations, ranging from -2.515 kcal/mol to -0.025 kcal/mol, while the stabilizing effect ranged from 0.023 to 1.975 kcal/mol.

The most destabilizing SNP across multiple populations was the CBR4 gene variant L70M, with a $\Delta\Delta$G value of -2.226 kcal/mol, while the most stabilizing SNP was the ABCC8 gene variant A1369S, with a $\Delta\Delta$G value of 1.289 kcal/mol. However, in the Indian population, the PCSK9 gene variant V460G exhibited the maximum destabilizing effect (-2.515 kcal/mol), while the ABCC8 gene variant E1274K had the maximum stabilizing effect (1.975 kcal/mol) across all populations.

The study's key discovery was that mutations in the MECR, ABCG8, and APOE genes resulted in protein destabilization, which was corroborated by changes in $\Delta\Delta$G values. Specifically, the F96L variant of the MECR gene (-1.01 kcal/mol), the Y54C variant of the ABCG8 gene (-1.277 kcal/mol), and the C130R variant of the APOE gene (-1.065 kcal/mol) exhibited destabilization with a decrease in $\Delta\Delta$G values.

Therefore, in order to investigate the potential mechanisms underlying the destabilization or stabilization caused by these variants, it would be advantageous to examine changes in the H-bond network, as hydrogen bonds are crucial for maintaining protein structure and even minor alterations in their patterns can significantly impact protein stability, folding, and function [58]. To accomplish this, we employed the HBPLUS



A) AFR

B) AMR

C) EAS

D) EUR

E) INDIGEN

F) SAS

Figure 3.14: Mutational effect over Protein stability and H-bond Network in RCT

program to analyze the loss and gain of hydrogen bonds in the H-bond network following the accumulation of the SNP.

Our results demonstrated that the variants either increased or decreased the number of hydrogen bonds formed. Across all populations except the Indian population, certain SNPs in ABCA1, APOE, ADRB1, and NPC1 genes led to an increase in the number of hydrogen bonds formed, while some SNPs in ABCG8 and APOH genes led to a decrease in the number of hydrogen bonds formed. The SNP in MVK-S52N caused a significant reduction in the number of native hydrogen bonds by 7, which may impact protein stability. In the Indian population, variants in ABCG8, ENPP2, and PRKAA2 genes exhibited the highest increase in the number of hydrogen bonds, whereas ABCG5 and APOH variants showed a decrease in the number of hydrogen bonds formed.

Furthermore, our findings indicated that the F96L variant of the MECR gene and the Y54C variant of the ABCG8 gene had no significant effect on the protein, while the C130R variant of the APOE gene, which plays a crucial role in the RCT pathway, resulted in the formation of an additional hydrogen bond.

These findings suggest that the C130R variant of the APOE gene may have a different impact on the RCT pathway compared to the native residue. The results provided that the F96L variant of the MECR gene and the Y54C variant of the ABCG8 gene were found to be indistinguishable from the native protein. It appears that the C130R variant of the APOE gene may exert a distinct effect on the RCT pathway relative to the wild-type residue. Conversely, our results demonstrated that the F96L variant of the MECR gene and the Y54C variant of the ABCG8 gene did not display any discernible functional differences compared to the native protein.

3.2.4 Exploring the Effects of Protein Mutations on Interactions and Pathways

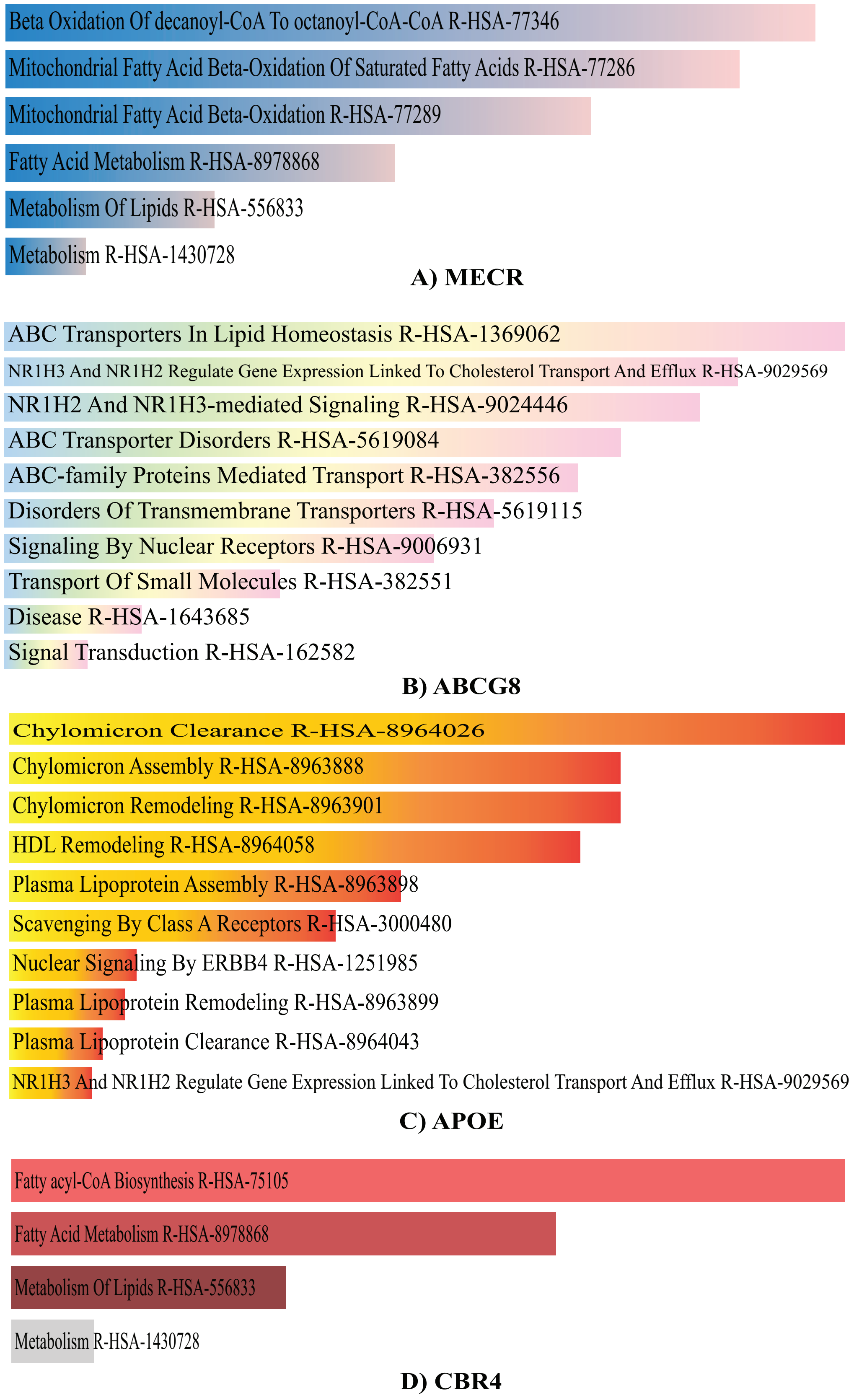
It is well-known that genes do not act in isolation and their interactions with other genes are crucial for proper cellular functioning. In this study, the focus was on the impact of non-synonymous variants on genes related to lipid metabolism. The analysis showed that these variants can significantly affect the interactions of target genes 3.16 with other genes in lipid-related pathways, such as fatty acid metabolism, lipid biosynthesis, and cholesterol metabolism 3.15. This suggests that non-synonymous single nucleotide polymorphisms (nsSNPs) in lipid-related genes can contribute to the development of lipid-related disorders such as dyslipidemia and atherosclerosis by altering the function of these genes and their interactions with other genes in the same pathways.

The Reactome database of pathways, using EnrichR, revealed that MECR is involved in various biological pathways related to fatty acid metabolism, such as beta-oxidation of decanoyl-CoA to octanoyl-CoA-CoA, mitochondrial fatty acid beta-oxidation of saturated fatty acids, fatty acid metabolism, and metabolism of lipids. Protein-protein interaction analysis using the STRING database showed that MECR is associated with other genes, such as ACAA2, ECHS1, FASN, HADHA, HADHB, and HTD2.

Similarly, ABCG8 is involved in various biological processes related to lipid homeostasis, such as ABC transporters in lipid homeostasis, NR1H3 and NR1H2 regulate gene expression linked to cholesterol transport and efflux, ABC transporter disorders, ABC-family proteins mediated transport, disorders of trans-membrane transporters, signaling by nuclear receptors, transport of small molecules, disease, and signal transduction. STRING database analysis also revealed that ABCG8 interacts with other genes, such as ABCA2, ABCB6, ABCD4, and ABCG5.

The APOE gene plays a significant role in several biological processes related to lipid metabolism, such as chylomicron clearance, chylomicron assembly, chylomicron remodeling, HDL remodeling, plasma lipoprotein assembly, scavenging by class A receptors, nuclear signaling by ERBB4, plasma lipoprotein remodeling, plasma lipoprotein clearance, and NR1H3 and NR1H2 regulate gene expression linked to cholesterol transport and efflux. STRING database analysis showed that the APOE gene interacts with other genes, such as APOA1, APOB, and several receptors involved in lipid metabolism, such as LDLR, LRP1, LRP8, and VLDLR.

In summary, the genes MECR, ABCG8, and APOE play a critical role in lipid metabolism and are linked with diverse biological pathways pertaining to fatty acid and cholesterol metabolism. The analysis of their interactions with other genes using the STRING database and EnrichR (Pathways) provided deeper insights into their function and downstream impact. The study findings demonstrate that these genes participate in intricate networks of gene interactions crucial for regular cellular function, and the impact of nsSNPs on these genes can have far-reaching consequences on various bio-



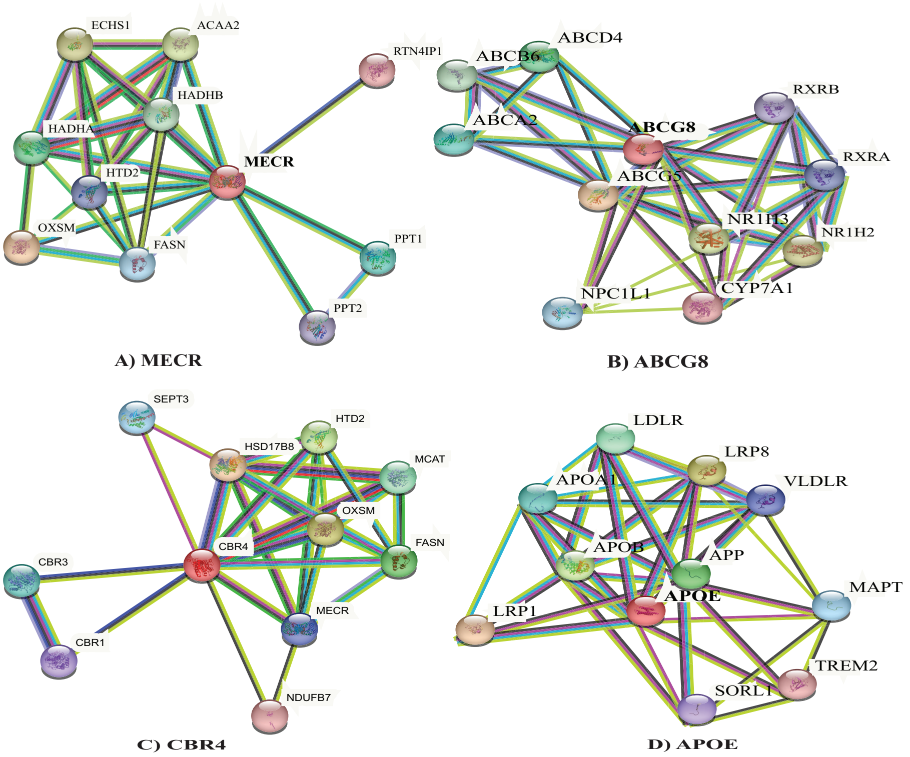
A) MECR

B) ABCG8

C) APOE

D) CBR4

Figure 3.15: Pathway Analysis



A) MECR

B) ABCG8

C) CBR4

D) APOE

Figure 3.16: Protein-Protein Interaction Analysis

logical pathways related to fatty acid and cholesterol metabolism. Overall, this research provides valuable insights into the impact of nsSNPs on the pathway that regulate the lipid balance in the body and emphasize the importance of understanding the genetic variations in diverse populations.

CONCLUSION

The thesis titled "Deciphering the Genetic Bias Underlying in Global Population for Lipid Homeostasis" delved into the impact of single nucleotide variants (SNVs) on the patterns of amino acid substitution profiles and the structural features of proteins in diverse populations. We analyzed 182 genes implicated in lipid metabolism, with a specific emphasis on their involvement in the reverse cholesterol pathway that is linked to cardiovascular disease. Our initial analysis identified 7,025 non-synonymous single nucleotide polymorphisms (nsSNPs) across the 182 genes involved in lipid homeostasis. Additionally, we identified 2,364 nsSNPs in 58 genes that are specifically associated with the reverse cholesterol pathway, using the 1kG and IndiGen sample datasets. Subsequently, we applied an allele frequency threshold of 10% to filter out the data and select the most common variants that contribute to the overall disease burden [19]. This resulted in a final set of 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) that were used in subsequent analysis. The likelihood of a missense mutation causing disease is influenced by various factors, including the mutation's location and nature, the protein affected, and the protein's physiological role. Our primary aim was to assess whether exploring the genetic basis of lipid homeostasis across diverse populations could enhance our understanding of the underlying mechanisms and aid in the development of more effective preventive measures and treatments. To identify the most common variants that contribute to the overall disease burden we had applied these two filters for filtering out the non-synonymous variants that should have minimum allele frequency of 10%. This resulted in a filtered set of variants that were used in the subsequent analysis. Afterwards we had conducted a thorough analysis of the frequencies of the amino acid substitutions for the protein variants associated with the lipid homeostasis and reverse cholesterol transport pathway to investigate the prevalent mutational pattern and to determine the similarities and differences in the amino acid substitution frequencies in each population present in 1kG and IndiGen sample data namely, African, American, South Asian, East Asian, European and Indian for all the reported nsSNps. Our comparative analysis for all the possible 20 standard amino acids present in the human genome revealed the distinct pattern the amino acid exchange profiles for different ethnicities. Due to dissimilarities in the amino acid exchange frequencies and mutability differences among diverse populations the matrix did not turned out to be balanced and several exchanges were found to be completely absent however certain others were found to be over-represented which indicated that the evolution of mutation may differ among different ethnic groups due to difference in ancestry. Across all the populations, in African population the most prevalent Phenylalanine to Leucine exchange was found which was contributed by MECR and LIPC gene while in the remaining populations the frequency for this exchange was found to be varied also within this population the Tryptophan to Cysteine substitution was found to population-specific with no existence in any other population. Except in the Indian population the Tyrosine to Cysteine exchange found to be prevalent across all the populations with the substitution frequency which was contributed by ABCG8 gene and in the Indian population the frequency for this exchange was found lower which was contributed by the DBI gene. In the Indian population, the most prevalent exchange was Iso-leucine to Serine. In addition to the most frequent exchanges, we also identified several exchanges that were common in all the populations and held significant importance despite occurred with differing frequencies. For instance, the Proline to Leucine exchange which was contributed by RBP1, FGFR4, TGFB1 and APOB genes and the Cysteine to Arginine exchange which was contributed by PEX2, PLD4, THADA and APOE genes. Similarly, we specifically performed an amino acid substitution analysis for the variants of the genes involved in the RCT pathway. The purpose of this analysis was to better understand the prevalence and potential impact of prevalent mutations identified in the substitution analysis of variants in lipid homeostasis pathway. We were interested in understanding how these variants may affect the function of the encoded proteins. We found that certain substitutions were consistently frequent across most populations, such as the F to L, Y to C and P to L substitutions. However, we observed that the frequency of these substitutions in the Indian population was relatively low with no occurrence of the Y to C exchange. Also, some population specific exchanges like W to C exchange were exclusively found in the African population with no reported occurrence in other populations. Additionally, we found that the C to R exchange was most prevalent in European and American populations, with a lesser occurrence frequency in the African population and no evidence in the EAS, SAS, and Indian populations. Thus, this observation suggests that these AA substitutions might be of greater importance in the RCT pathway. The results indicate that the African population might had undergone more extensive genetic divergence compared to other populations leading to the accumulation of distinct AA substitutions that were found to be unique in the African population. To further examine the potential genetic differences between populations, a statistical comparison of amino acid exchange frequencies was conducted. Since, amino acid substitution basically involves comparing frequencies across multiple populations, the statistical test was performed to determine whether the observed differences in substitution frequencies were statistically significant or due to chance. To analyze the count variable data and to compare the substitution frequencies between population pairs, a non-parametric Mann Whitney U-test was used. Thus, the results revealed that the African population had a distinct pattern of amino acid substitution frequency compared to other populations, indicating unique evolutionary history. The unique amino acid substitutions found in the African population was supported by our analysis, which determined that certain amino acids are more prone to mutations in these metabolic pathways. To accomplish this, we calculated the mutabilities of all standard amino acids across diverse populations. Overall, we observed that different populations exhibit different levels of mutability for the 20 amino acids in both the pathways where the Valine was found to be most mutable in AMR, EAS, EUR, and SAS population while Arginine in AFR and AMR population. Also, Threonine was found to have the highest mutability in EUR and SAS population while Indian population exhibit unique mutable amino acids alanine and serine. Furthermore, the codon usage bias was observed in certain populations with lower mutation frequencies for the amino acids coded by a greater number of codons compared to those coded by fewer codons. This might be due to the distinct genetic histories or environmental conditions that could result in difference in mutation rate. Consequently, these amino acid substitutions can induce alterations in the chemical characteristics of proteins which could impact their structure and function. To elucidate the nature of whether these amino acid substitutions involve inter-class or intra-class conversions, the chemical shift was analyzed. Based on our chemical shift analysis results across six populations more inter-class conversions were observed than the intra-class conversions in both the pathways which was evidenced by the high number of non-zero values off the diagonal compared to on the diagonal. The intra-class class conversion was found to be prevalent among Aliphatic to Aliphatic class in all the populations. During our amino acid substitution analysis, we identified prevalent common substitutions across all populations and wanted to observe their chemical shift. Our findings indicated that all of these substitutions, involved in inter-class conversions, had the potential to cause significant changes in protein structure and function. To determine the location of these variants and their potential impact, we classified them into pre-domain, within domain, and post-domain regions based on their position. This classification was necessary because variants in the domain region are more likely to affect protein structure compared to those located nearby. The proportion of variants in each region (pre-domain, post-domain, and within domain) is represented by the angle of each bar in the graph. Our findings suggest a bias towards within domain and post-domain regions as there were fewer variants in the pre-domain region. Among the Indian population, the majority of the variants (91) were found in the domain region, 136 were found in the post domain region while only 2 were located in the pre-domain region. The African population had the maximum number of variants (155) in the post-domain region. Additionally, all of the prevalent exchanges we identified were located within the domain region for these genes. After comprehending the impact of variants at the sequence level analysis it becomes essential to have a more profound understanding of the effect of the variants at the structural level. So we had conducted the structural level analysis to have a deeper insight into the implications of the variants over protein structure. To investigate the impact of the identified variants we applied several filters to the sequence analysis processed data, including the nsSNPs and minimum allele frequency filter of 10%, availability of at least one protein crystal structure, sequence coverage of at least 70% and SNV coverage to the crystal structure. After applying these filters; we obtained a subset of data consisting of 31 variants in 20 genes for African population, 22 variants in 15 genes for American population, 24 variants in 17 genes for East Asian population, 22 variants in 16 genes for European population 53 variants in 28 genes for Indian population and 22 variants in 15 genes for South Asian population. The limitation of our study was that our data was reduced due to the unavailability of protein crystal structure, despite the wide availability of protein sequence information. Moreover, the few available protein structures did not meet our filter conditions, further limit our analysis. To conduct our analysis, we generated variant models using Modeller, which were then energy minimized using GROMACS. Loop refinement was carried out using Modrefiner. All subsequent analysis was performed using these variant models.

Analyzing the secondary structure of the non-synonymous variants can provide insight into how these mutations affect the protein's conformation and stability so we used the Dictionary of secondary structure of proteins. Interestingly, our analysis revealed that across all the populations, a higher percentage of SNPs mapped to the other regions (including; loop, coils) than alpha-helix or beta sheets. However, certain non-synonymous variants like CBR4-L70M and THOC5-V525I resulted in changes in secondary structure across all populations except for the Indian population. Specifically, in the native protein, the affected amino acid residue was found to be in a turn and strand conformation, while after mutation accumulation, the mutated residue at the same position in the reference site was found in an alpha-helix and other regions of the protein. Overall, approximately 9 gene variants that resulted in alterations in the conformation of their secondary structure in the Indian population highlights the genetic diversity of this population. One of these 9 variants is CBR4-L70Q where the substitution occurs at the same reference site as in CBR4-L70M but the reference amino acid is substituted by different mutant amino acids in the Indian population. For the remaining gene variants across all the populations the secondary structure remained unchanged. Besides this, we looked at our prevalent exchanges from sequence analysis, where the F to L substitution found at position 96 of MECR gene which have a role in pathway for the catabolism of fatty acids in the mitochondria, the Y to C substitution was found at position 54 of ABCG8 gene which is thought to play a role in the delivery of cholesterol to the liver by facilitating the transport of cholesterol from the circulation into the liver cells. It is a member of the ABC transporter family, and is expressed in the liver, small intestine, and other tissues and the C to R substitution found at position 130 of APOE gene which has a role in the transport and metabolism of cholesterol and other lipids in the body. It is a component of HDL and LDL, and is also involved in the formation and metabolism of triglycerides.This gene found to have role in the second phase of RCT which involves the maturation of discoidal HDL (dHDL) into spherical HDL (sHDL). This process involves the addition of more proteins, including APOA1 and APOE, to the dHDL particle, which increases its size and stability. It's also noteworthy that the prevalent exchanges did not cause any alteration in the secondary structure conformation of the affected amino acid residues. While it had been observed that certain variants result in a change in the secondary structure of the protein and may be destabilizing, the extent of this destabilization remains to be determined. Although several other variants did not result in any detectable change in the secondary structure conformation of the protein, they still could have significant destabilizing or stabilizing effects. Therefore, to carefully evaluate the impact of these variants on protein stability we analyzed the changes in the delta-delta G values calculated using Dynamut2 API in Kcal/mol. Our results revealed that the majority of the variants had a destabilizing effect over the protein structure with only few having a stabilizing effect. The extent of destabilization varied across six populations ranging from -2.515 kcal/mol to -0.025 kcal/mol while the magnitude of stabilization varied from 0.023 kcal/mol to 1.975 kcal/mol. Across all the populations the CBR4-L70M and the CBR4-L70Q variant caused changes in their secondary structure were found to have the significant destabilizing effect on the protein while the THOC5-V525I variant also found to have minor destabilizing effect. In addition to variants that were found to cause destabilization to the protein, certain variants were also observed to have a stabilizing effect like ABCC8-A1369S. The key discovery was that the mutation in MECR, ABCG8 and APOE gene resulted in destabilization which was corroborated by the changes in the delta delta G values. Among these three variants the ABCG8 variant was found to have maximum destabilizing effect while the MECR variant with minor destabilization and the APOE gene variant showed moderate destabilizing effect. Although several variants had been identified to cause significant destabilization to the protein, it becomes essential to understand the underlying mechanisms that may contribute to this destabilizing effect. One potential factor that can influence protein stability is the hydrogen bonding network, which can be disrupted by changes in the amino acid sequence. So, we had analyzed the hydrogen bonding network within the protein to identify any changes that might be associated with destabilization so we used the HBPLUS program to analyze the loss and gain of hydrogen bonds in the H-bond network. We had observed that the variants had a significant impact on the hydrogen bond network. Specifically, we observed that some variants resulted in an increase in the number of hydrogen bonds, while others led to a decrease. Interestingly, the CBR4-L70M and THOC5-V525I variants had the same number of hydrogen bonds in both the native and mutant protein structures. This suggests that the variant may not directly impact the stability of the protein through changes in the hydrogen bonding network while the CBR4-L70Q variant had an additional hydrogen bond in the mutant protein structure compared to the native protein. The MECR-F96L variant and ABCG8-Y54C variant had no significant effect on the protein suggests that these variants are likely to be benign or neutral with respect to their impact on protein function. while the C130R variant of APOE gene resulted in the formation of addition of hydrogen bond suggests that this variant could have a significant impact on protein structure and function. In this study, we classified the solvent accessibility of amino acid residues based on a threshold of <=30% for buried, >=70% for exposed, & >30% & <70% for either buried or exposed. This classification helped us to determine whether the amino acid residues were exposed or buried within the protein structure. After analyzing the variants, we observed that the maximum number of variants were located in buried residues, while only a few were in exposed residues. We used this information to further analyze the impact of mutations on the protein structure and function. We found that in their native protein state, all affected amino acid residues in the CBR4-L70M, THOC5-V5252I, MECR-F96L, CBR4-L70Q, and ABCG8-Y54C variants were buried. However, as mutations accumulated, some of these residues became more exposed on the protein's surface. This finding suggests that mutations can have a significant impact on the solvent accessibility of amino acid residues and thus on the protein structure and function. To analyze the functional implications of the identified variants, we performed protein-protein interaction (PPI) analysis. Our analysis revealed that these genes have significant roles in lipid metabolism and are associated with various biological pathways related to fatty acid and cholesterol metabolism. Interestingly, the PPI analysis showed that the identified variants may impair the protein's ability to interact with other proteins that are critical for the proper functioning of lipid metabolism. These changes in protein-protein interactions could lead to downstream effects on lipid metabolism, including impaired fatty acid and cholesterol metabolism, altered sterol transport and excretion, and compromised bile acid synthesis that may contribute to the development and progression of metabolic disorders such as hypercholesterolemia, atherosclerosis, and non-alcoholic fatty liver disease. Additionally, our findings suggest that these genes interact with other genes and are involved in complex networks of gene interactions, which are vital for maintaining cellular function. Understanding how nsSNP accumulation in a gene affects pathways is crucial for gaining insights into the underlying biological processes and disease susceptibility. By identifying the pathways impacted by these genetic variations, we can develop targeted interventions to correct or modulate the effects. In our analysis, we found that our genes belong to various pathways, including HDL remodeling, LDL clearance, and others. We also observed that some genes share common pathways, such as MECR and CBR4, which suggests their involvement in similar biological processes and potential synergistic effects on these pathways. On the other hand, ABCG8 and APOE share common pathways, indicating complex interaction and impact on the RCT pathway.

This approach may help in identifying lifestyle and environmental factors that modify the risk of lipid homeostasis-related diseases and disorders, and inform interventions to prevent or alleviate these conditions. The novelty of our study lies in its broader scope as prior investigations have typically concentrated on a limited number of populations, impeding a comprehensive understanding of the global genetic architecture.

Future Goals

1) Determination of binding site: It could be important in understanding the genetic basis of lipid homeostasis in the global population. This is because many proteins involved in lipid metabolism have specific binding sites that allow them to interact with other molecules, such as lipids or other proteins. By identifying these binding sites and understanding how they contribute to the protein's function, researchers can better understand the role that the protein plays in the body and how it contributes to the overall regulation of lipid homeostasis.

2) Investigating the functional consequences of the identified variants: Understanding how the identified genetic variations affect protein structure and function can provide insights into potential targets for therapeutic interventions in the regulation of lipid balance and RCT pathway.

3) Expanding the study to include larger sample sizes and more diverse populations: Expanding the study to include more diverse populations with larger sample sizes can provide a more comprehensive understanding of the genetic basis of lipid metabolism and RCT pathway regulation and how it varies across different populations.

4) Developing personalized treatments for lipid-related diseases: By identifying genetic variations that contribute to lipid-related disease risk, personalized treatments could be developed that target specific variants or pathways involved in lipid metabolism and RCT pathway regulation. This approach could potentially improve treatment outcomes and reduce the burden of lipid-related diseases.

REFERENCES

[1] N. A. Abumrad and N. O. Davidson, “Role of the gut in lipid homeostasis,” *Physiological reviews*, vol. 92, no. 3, pp. 1061–1085, 2012.

[2] E. Agmon and B. R. Stockwell, “Lipid homeostasis and regulated cell death,” *Current opinion in chemical biology*, vol. 39, pp. 83–89, Aug. 2017.

[3] G. Van Meer, D. R. Voelker, and G. W. Feigenson, “Membrane lipids: where they are and how they behave,” *Nature reviews Molecular cell biology*, vol. 9, no. 2, pp. 112–124, 2008.

[4] “Lipid homeostasis Related Genes And Facts | Bosterbio,” https://www.bosterbio.com/pathways/lipid-homeostasis.

[5] “CHOLESTEROL METABOLISM.docx - CHOLES TEROL Despite having a BAD reputation as a high-risk factor for cardiovascular diseases, cholesterol is an | Course Hero,” https://www.coursehero.com/file/114269966/CHOLESTEROL-METABOLISMdocx/.

[6] “In the intestine the dietary fats are hydrolysed by? Explained by FAQ Blog,” https://hep.dcmusic.ca/in-the-intestine-the-dietary-fats-are-hydrolysed-by.

[7] H. R. Waterham, “Defects of cholesterol biosynthesis,” *FEBS letters*, vol. 580, no. 23, pp. 5442–5449, 2006.

[8] D. W. Russell, “Cholesterol biosynthesis and metabolism,” *Cardiovascular drugs and therapy*, vol. 6, no. 2, pp. 103–110, 1992.

[9] C. J. Fielding and P. E. Fielding, “Molecular physiology of reverse cholesterol transport.” *Journal of lipid research*, vol. 36, no. 2, pp. 211–228, 1995.

[10] R. S. Rosenson, H. B. Brewer Jr, W. S. Davidson, Z. A. Fayad, V. Fuster, J. Goldstein, M. Hellerstein, X.-C. Jiang, M. C. Phillips, D. J. Rader *et al.*, “Cholesterol efflux and atheroprotection: advancing the concept of reverse cholesterol transport,” *Circulation*, vol. 125, no. 15, pp. 1905–1919, 2012.

[11] "Lipid Storage Diseases Fact Sheet | National Institute of Neurological Disorders and Stroke," https://www.ninds.nih.gov/lipid-storage-diseases-fact-sheet.

[12] Z. Zhang, M. A. Miteva, L. Wang, and E. Alexov, "Analyzing effects of naturally occurring missense mutations," *Computational and mathematical methods in medicine*, vol. 2012, 2012.

[13] "Genetic Mutation | Learn Science at Scitable," http://www.nature.com/scitable/topicpage/genetic-mutation-441.

[14] P. Prakrithi, K. Singhal, D. Sharma, A. Jain, R. C. Bhoyar, M. Imran, V. Senthilvel, M. K. Divakar, A. Mishra, V. Scaria, S. Sivasubbu, and M. Mukerji, "An Alu insertion map of the Indian population: Identification and analysis in 1021 genomes of the IndiGen project," *NAR Genomics and Bioinformatics*, vol. 4, no. 1, p. lqac009, Mar. 2022.

[15] Avantika, "What Are Mutations?Definition, Causes and Effects of Mutations," https://byjus.com/biology/mutation-genetic-change/.

[16] "Point Mutation," https://www.vedantu.com/biology/point-mutation.

[17] T. A. P. de Beer, R. A. Laskowski, S. L. Parks, B. Sipos, N. Goldman, and J. M. Thornton, "Amino Acid Changes in Disease-Associated Variants Differ Radically from Variants Observed in the 1000 Genomes Project Dataset," *PLoS Computational Biology*, vol. 9, no. 12, p. e1003382, Dec. 2013.

[18] M. Soskine and D. S. Tawfik, "Mutational effects and the evolution of new protein functions," *Nature Reviews Genetics*, vol. 11, no. 8, pp. 572–582, 2010.

[19] G. Panda, N. Mishra, D. Sharma, R. Kutum, R. C. Bhoyar, A. Jain, M. Imran, V. Senthilvel, M. K. Divakar, A. Mishra, P. Garg, P. Banerjee, S. Sivasubbu, V. Scaria, and A. Ray, "Comprehensive assessment of Indian variations in the druggable kinome landscape highlights distinct insights at the sequence, structure and pharmacogenomic stratum," p. 2021.05.23.445314, Nov. 2021.

[20] A. J. Marian, "Molecular genetic studies of complex phenotypes," *Translational Research*, vol. 159, no. 2, pp. 64–79, 2012.

[21] Z. Pan and S. Xu, "Population genomics of East Asian ethnic groups," *Hereditas*, vol. 157, no. 1, p. 49, Dec. 2020.

[22] G. Abrusán and J. A. Marsh, "Alpha Helices Are More Robust to Mutations than Beta Strands," *PLOS Computational Biology*, vol. 12, no. 12, p. e1005242, Dec. 2016.

[23] C. H. Rodrigues, D. E. Pires, and D. B. Ascher, "DynaMut2: Assessing changes in stability and flexibility upon single and multiple point missense mutations," *Protein Science*, vol. 30, no. 1, pp. 60–69, 2021.

[24] C. Dempsey, "Geography of Ancestry," Aug. 2012.

[25] C. Zhang, Y. Gao, Z. Ning, Y. Lu, X. Zhang, J. Liu, B. Xie, Z. Xue, X. Wang, K. Yuan, X. Ge, Y. Pan, C. Liu, L. Tian, Y. Wang, D. Lu, B.-P. Hoh, and S. Xu, "PGG.SNV: Understanding the evolutionary and medical implications of human single nucleotide variations in diverse populations," *Genome Biology*, vol. 20, no. 1, p. 215, Oct. 2019.

[26] S. Belsare, M. Levy-Sakin, Y. Mostovoy, S. Durinck, S. Chaudhuri, M. Xiao, A. S. Peterson, P.-Y. Kwok, S. Seshagiri, and J. D. Wall, "Evaluating the quality of the 1000 genomes project data," *BMC genomics*, vol. 20, no. 1, pp. 1–14, 2019.

[27] . G. P. Consortium *et al.*, "A global reference for human genetic variation," *Nature*, vol. 526, no. 7571, p. 68, 2015.

[28] K. J. Karczewski, L. C. Francioli, G. Tiao, B. B. Cummings, J. Alföldi, Q. Wang, R. L. Collins, K. M. Laricchia, A. Ganna, D. P. Birnbaum *et al.*, "The mutational constraint spectrum quantified from variation in 141,456 humans," *Nature*, vol. 581, no. 7809, pp. 434–443, 2020.

[29] G. Sirugo, S. M. Williams, and S. A. Tishkoff, "The missing diversity in human genetic studies," *Cell*, vol. 177, no. 1, pp. 26–31, 2019.

[30] I. G. V. C. .-.-. .-.-. skb@ igib. res. in, "The indian genome variation database (igvdb): a project overview," *Human genetics*, vol. 118, pp. 1–11, 2005.

[31] A. Jain, R. C. Bhoyar, K. Pandhare, A. Mishra, D. Sharma, M. Imran, V. Senthivel, M. K. Divakar, M. Rophina, B. Jolly, A. Batra, S. Sharma, S. Siwach, A. G.

Jadhao, N. V. Palande, G. N. Jha, N. Ashrafi, P. K. Mishra, V. A. K., S. Jain, D. Dash, N. S. Kumar, A. Vanlallawma, R. J. Sarma, L. Chhakchhuak, S. Kalyanaraman, R. Mahadevan, S. Kandasamy, P. B. M., R. E. Rajagopal, E. R. J., N. D. P., A. Bajaj, V. Gupta, S. Mathew, S. Goswami, M. Mangla, S. Prakash, K. Joshi, Meyakumla, S. S., D. Gajjar, R. Soraisham, R. Yadav, Y. S. Devi, A. Gupta, M. Mukerji, S. Ramalingam, B. B. K., V. Scaria, and S. Sivasubbu, "IndiGenomes: A comprehensive resource of genetic variants from over 1000 Indian genomes," *Nucleic Acids Research*, p. gkaa923, Oct. 2020.

[32] K. Wang, M. Li, and H. Hakonarson, "ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data," *Nucleic Acids Research*, vol. 38, no. 16, p. e164, Sep. 2010.

[33] G. Drin, "Topological regulation of lipid balance in cells," *Annual review of biochemistry*, vol. 83, pp. 51–77, 2014.

[34] H. Wang, M. Lei, R.-c. Hsia, and C. Sztalryd, "Chapter 8 - Analysis of Lipid Droplets in Cardiac Muscle," in *Methods in Cell Biology*, ser. Lipid Droplets, H. Yang and P. Li, Eds. Academic Press, Jan. 2013, vol. 116, pp. 129–149.

[35] D. Smedley, S. Haider, B. Ballester, R. Holland, D. London, G. Thorisson, and A. Kasprzyk, "BioMart–biological queries made easy," *BMC genomics*, vol. 10, p. 22, Jan. 2009.

[36] W. Li, J. Wypych, and R. J. Duff, "Improved sequence variant analysis strategy by automated false positive removal," in *MAbs*, vol. 9, no. 6. Taylor & Francis, 2017, pp. 978–984.

[37] H. Venselaar, T. A. Te Beek, R. K. Kuipers, M. L. Hekkelman, and G. Vriend, "Protein structure analysis of mutations causing inheritable diseases. an e-science approach with life scientist friendly interfaces," *BMC bioinformatics*, vol. 11, no. 1, pp. 1–10, 2010.

[38] "Filter-based Annotation - ANNOVAR Documentation," https://annovar.openbioinformatics.org/en/latest/user-guide/filter/.

[39] Z. Zhang, M. A. Miteva, L. Wang, and E. Alexov, "Analyzing Effects of Natu-

rally Occurring Missense Mutations," *Computational and Mathematical Methods in Medicine*, vol. 2012, p. 805827, 2012.

[40] "Protein folding and misfolding | Nature," https://www.nature.com/articles/nature02261.

[41] P. M. Silverman, G. S. Harell, and M. Korobkin, "Computed tomography of the abnormal pericardium," *AJR American journal of roentgenology*, vol. 140, no. 6, pp. 1125–1129, Jun. 1983.

[42] A. Šali and T. L. Blundell, "Comparative Protein Modelling by Satisfaction of Spatial Restraints," *Journal of Molecular Biology*, vol. 234, no. 3, pp. 779–815, Dec. 1993.

[43] "GROMACS," *Wikipedia*, Sep. 2022.

[44] "ModRefiner: High-resolution Protein Structure Refinement and Relaxation by Energy Minimization," https://zhanggroup.org/ModRefiner/.

[45] "DSSP," https://swift.cmbi.umcn.nl/gv/dssp/DSSP_3.html.

[46] "Naccess - Mathematical software - swMATH," https://www.swmath.org/software/12485.

[47] "HBPLUS home page," https://www.ebi.ac.uk/thornton-srv/software/HBPLUS/.

[48] M. M. Gromiha, "Chapter 6 - Protein Stability," in *Protein Bioinformatics*, M. M. Gromiha, Ed. Singapore: Academic Press, Jan. 2010, pp. 209–245.

[49] 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, Jul. 2018.

[50] S. Gravel, B. M. Henn, R. N. Gutenkunst, A. R. Indap, G. T. Marth, A. G. Clark, F. Yu, R. A. Gibbs, The 1000 Genomes Project, C. D. Bustamante, D. L. Altshuler, R. M. Durbin, G. R. Abecasis, D. R. Bentley, A. Chakravarti, A. G. Clark, F. S. Collins, F. M. De La Vega, P. Donnelly, M. Egholm, P. Flicek, S. B. Gabriel, R. A. Gibbs, B. M. Knoppers, E. S. Lander, H. Lehrach, E. R. Mardis, G. A. McVean, D. A. Nickerson, L. Peltonen, A. J. Schafer, S. T. Sherry, J. Wang, R. K. Wilson, R. A. Gibbs, D. Deiros, M. Metzker, D. Muzny, J. Reid, D. Wheeler,

J. Wang, J. Li, M. Jian, G. Li, R. Li, H. Liang, G. Tian, B. Wang, J. Wang, W. Wang, H. Yang, X. Zhang, H. Zheng, E. S. Lander, D. L. Altshuler, L. Ambrogio, T. Bloom, K. Cibulskis, T. J. Fennell, S. B. Gabriel, D. B. Jaffe, E. Shefler, C. L. Sougnez, D. R. Bentley, N. Gormley, S. Humphray, Z. Kingsbury, P. Koko-Gonzales, J. Stone, K. J. McKernan, G. L. Costa, J. K. Ichikawa, C. C. Lee, R. Sudbrak, H. Lehrach, T. A. Borodina, A. Dahl, A. N. Davydov, P. Marquardt, F. Mertes, W. Nietfeld, P. Rosenstiel, S. Schreiber, A. V. Soldatov, B. Timmermann, M. Tolzmann, M. Egholm, J. Affourtit, D. Ashworth, S. Attiya, M. Bachorski, E. Buglione, A. Burke, A. Caprio, C. Celone, S. Clark, D. Conners, B. Desany, L. Gu, L. Guccione, K. Kao, A. Kebbel, J. Knowlton, M. Labrecque, L. McDade, C. Mealmaker, M. Minderman, A. Nawrocki, F. Niazi, K. Pareja, R. Ramenani, D. Riches, W. Song, C. Turcotte, S. Wang, E. R. Mardis, R. K. Wilson, D. Dooling, L. Fulton, R. Fulton, G. Weinstock, R. M. Durbin, J. Burton, D. M. Carter, C. Churcher, A. Coffey, A. Cox, A. Palotie, M. Quail, T. Skelly, J. Stalker, H. P. Swerdlow, D. Turner, A. De Witte, S. Giles, R. A. Gibbs, D. Wheeler, M. Bainbridge, D. Challis, A. Sabo, F. Yu, J. Yu, J. Wang, X. Fang, X. Guo, R. Li, Y. Li, R. Luo, S. Tai, H. Wu, H. Zheng, X. Zheng, Y. Zhou, G. Li, J. Wang, H. Yang, G. T. Marth, E. P. Garrison, W. Huang, A. Indap, D. Kural, W.-P. Lee, W. F. Leong, A. R. Quinlan, C. Stewart, M. P. Stromberg, A. N. Ward, J. Wu, C. Lee, R. E. Mills, X. Shi, M. J. Daly, M. A. DePristo, D. L. Altshuler, A. D. Ball, E. Banks, T. Bloom, B. L. Browning, K. Cibulskis, T. J. Fennell, K. V. Garimella, S. R. Grossman, R. E. Handsaker, M. Hanna, C. Hartl, D. B. Jaffe, A. M. Kernytsky, J. M. Korn, H. Li, J. R. Maguire, S. A. McCarroll, A. McKenna, J. C. Nemesh, A. A. Philippakis, R. E. Poplin, A. Price, M. A. Rivas, P. C. Sabeti, S. F. Schaffner, E. Shefler, I. A. Shlyakhter, D. N. Cooper, E. V. Ball, M. Mort, A. D. Phillips, P. D. Stenson, J. Sebat, V. Makarov, K. Ye, S. C. Yoon, C. D. Bustamante, A. G. Clark, A. Boyko, J. Degenhardt, S. Gravel, R. N. Gutenkunst, M. Kaganovich, A. Keinan, P. Lacroute, X. Ma, A. Reynolds, L. Clarke, P. Flicek, F. Cunningham, J. Herrero, S. Keenen, E. Kulesha, R. Leinonen, W. M. McLaren, R. Radhakrishnan, R. E. Smith, V. Zalunin, X. Zheng-Bradley, J. O. Korbel, A. M. Stütz, S. Humphray, M. Bauer, R. K. Cheetham, T. Cox, M. Eberle, T. James, S. Kahn, L. Murray, A. Chakravarti, K. Ye, F. M. De La Vega, Y. Fu, F. C. L. Hyland, J. M. Manning, S. F. McLaughlin, H. E. Peckham, O. Sakarya, Y. A. Sun, E. F.

Tsung, M. A. Batzer, M. K. Konkel, J. A. Walker, R. Sudbrak, M. W. Albrecht, V. S. Amstislavskiy, R. Herwig, D. V. Parkhomchuk, S. T. Sherry, R. Agarwala, H. M. Khouri, A. O. Morgulis, J. E. Paschall, L. D. Phan, K. E. Rotmistrovsky, R. D. Sanders, M. F. Shumway, C. Xiao, G. A. McVean, A. Auton, Z. Iqbal, G. Lunter, J. L. Marchini, L. Moutsianas, S. Myers, A. Tumian, B. Desany, J. Knight, R. Winer, D. W. Craig, S. M. Beckstrom-Sternberg, A. Christoforides, A. A. Kurdoglu, J. V. Pearson, S. A. Sinari, W. D. Tembe, D. Haussler, A. S. Hinrichs, S. J. Katzman, A. Kern, R. M. Kuhn, M. Przeworski, R. D. Hernandez, B. Howie, J. L. Kelley, S. C. Melton, G. R. Abecasis, Y. Li, P. Anderson, T. Blackwell, W. Chen, W. O. Cookson, J. Ding, H. M. Kang, M. Lathrop, L. Liang, M. F. Moffatt, P. Scheet, C. Sidore, M. Snyder, X. Zhan, S. Zöllner, P. Awadalla, F. Casals, Y. Idaghdour, J. Keebler, E. A. Stone, M. Zilversmit, L. Jorde, J. Xing, E. E. Eichler, G. Aksay, C. Alkan, I. Hajirasouliha, F. Hormozdiari, J. M. Kidd, S. C. Sahinalp, P. H. Sudmant, E. R. Mardis, K. Chen, A. Chinwalla, L. Ding, D. C. Koboldt, M. D. McLellan, D. Dooling, G. Weinstock, J. W. Wallis, M. C. Wendl, Q. Zhang, R. M. Durbin, C. A. Albers, Q. Ayub, S. Balasubramaniam, J. C. Barrett, D. M. Carter, Y. Chen, D. F. Conrad, P. Danecek, E. T. Dermitzakis, M. Hu, N. Huang, M. E. Hurles, H. Jin, L. Jostins, T. M. Keane, S. Q. Le, S. Lindsay, Q. Long, D. G. MacArthur, S. B. Montgomery, L. Parts, J. Stalker, C. Tyler-Smith, K. Walter, Y. Zhang, M. B. Gerstein, M. Snyder, A. Abyzov, S. Balasubramanian, R. Bjornson, J. Du, F. Grubert, L. Habegger, R. Haraksingh, J. Jee, E. Khurana, H. Y. K. Lam, J. Leng, X. J. Mu, A. E. Urban, Z. Zhang, Y. Li, R. Luo, G. T. Marth, E. P. Garrison, D. Kural, A. R. Quinlan, C. Stewart, M. P. Stromberg, A. N. Ward, J. Wu, C. Lee, R. E. Mills, X. Shi, S. A. McCarroll, E. Banks, M. A. DePristo, R. E. Handsaker, C. Hartl, J. M. Korn, H. Li, J. C. Nemesh, J. Sebat, V. Makarov, K. Ye, S. C. Yoon, J. Degenhardt, M. Kaganovich, L. Clarke, R. E. Smith, X. Zheng-Bradley, J. O. Korbel, S. Humphray, R. K. Cheetham, M. Eberle, S. Kahn, L. Murray, K. Ye, F. M. De La Vega, Y. Fu, H. E. Peckham, Y. A. Sun, M. A. Batzer, M. K. Konkel, J. A. Walker, C. Xiao, Z. Iqbal, B. Desany, T. Blackwell, M. Snyder, J. Xing, E. E. Eichler, G. Aksay, C. Alkan, I. Hajirasouliha, F. Hormozdiari, J. M. Kidd, K. Chen, A. Chinwalla, L. Ding, M. D. McLellan, J. W. Wallis, M. E. Hurles, D. F. Conrad, K. Walter, Y. Zhang, M. B. Gerstein, M. Snyder, A. Abyzov, J. Du, F. Grubert, R. Haraksingh, J. Jee, E. Khurana,

H. Y. K. Lam, J. Leng, X. J. Mu, A. E. Urban, Z. Zhang, R. A. Gibbs, M. Bainbridge, D. Challis, C. Coafra, H. Dinh, C. Kovar, S. Lee, D. Muzny, L. Nazareth, J. Reid, A. Sabo, F. Yu, J. Yu, G. T. Marth, E. P. Garrison, A. Indap, W. F. Leong, A. R. Quinlan, C. Stewart, A. N. Ward, J. Wu, K. Cibulskis, T. J. Fennell, S. B. Gabriel, K. V. Garimella, C. Hartl, E. Shefler, C. L. Sougnez, J. Wilkinson, A. G. Clark, S. Gravel, F. Grubert, L. Clarke, P. Flicek, R. E. Smith, X. Zheng-Bradley, S. T. Sherry, H. M. Khouri, J. E. Paschall, M. F. Shumway, C. Xiao, G. A. McVean, S. J. Katzman, G. R. Abecasis, T. Blackwell, E. R. Mardis, D. Dooling, L. Fulton, R. Fulton, D. C. Koboldt, R. M. Durbin, S. Balasubramaniam, A. Coffey, T. M. Keane, D. G. MacArthur, A. Palotie, C. Scott, J. Stalker, C. Tyler-Smith, M. B. Gerstein, S. Balasubramanian, A. Chakravarti, B. M. Knoppers, G. R. Abecasis, C. D. Bustamante, N. Gharani, R. A. Gibbs, L. Jorde, J. S. Kaye, A. Kent, T. Li, A. L. McGuire, G. A. McVean, P. N. Ossorio, C. N. Rotimi, Y. Su, L. H. Toji, C. TylerSmith, L. D. Brooks, A. L. Felsenfeld, J. E. McEwen, A. Abdallah, C. R. Juenger, N. C. Clemm, F. S. Collins, A. Duncanson, E. D. Green, M. S. Guyer, J. L. Peterson, A. J. Schafer, G. R. Abecasis, D. L. Altshuler, A. Auton, L. D. Brooks, R. M. Durbin, R. A. Gibbs, M. E. Hurles, and G. A. McVean, “Demographic history and rare allele sharing among human populations,” *Proceedings of the National Academy of Sciences*, vol. 108, no. 29, pp. 11 983–11 988, Jul. 2011.

[51] Z. Peng, H. Zaher, and Y. Ben-Shahar, “Natural selection on gene-specific codon usage bias is common across eukaryotes,” p. 292938, May 2018.

[52] M. Soskine and D. S. Tawfik, “Mutational effects and the evolution of new protein functions,” *Nature Reviews Genetics*, vol. 11, no. 8, pp. 572–582, Aug. 2010.

[53] V. M. Ingram, “A specific chemical difference between the globins of normal human and sickle-cell anaemia haemoglobin,” *Nature*, vol. 178, no. 4537, pp. 792–794, Oct. 1956.

[54] A. C. Joerger and A. R. Fersht, “The tumor suppressor p53: From structures to drug discovery,” *Cold Spring Harbor Perspectives in Biology*, vol. 2, no. 6, p. a000919, Jun. 2010.

[55] R. W. Mahley and Y. Huang, “Apolipoprotein e sets the stage: Response to injury triggers neuropathology,” *Neuron*, vol. 76, no. 5, pp. 871–885, Dec. 2012.

[56] S. H. Cheng, R. J. Gregory, J. Marshall, S. Paul, D. W. Souza, G. A. White, C. R. O'Riordan, and A. E. Smith, "Defective intracellular transport and processing of CFTR is the molecular basis of most cystic fibrosis," *Cell*, vol. 63, no. 4, pp. 827–834, Nov. 1990.

[57] E. Feyfant, A. Sali, and A. Fiser, "Modeling mutations in protein structures," *Protein Science*, vol. 16, no. 9, pp. 2030–2041, 2007.

[58] R. M. Razban and E. I. Shakhnovich, "Effects of Single Mutations on Protein Stability Are Gaussian Distributed," *Biophysical Journal*, vol. 118, no. 12, pp. 2872–2878, Jun. 2020.