Universiteit Leiden

Master Computer Science

Exploring Differences in ACMG Pathogenic Variants Among Diverse Populations through Comparative Analysis

Name:	Aleksandra Baumgart
Student ID:	3736245
Date:	24/07/2024
Specialisation:	Bioinformatics
1st supervisor:	Dr Katy Wolstencroft
2nd supervisor:	Dr Julij Selb

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Contents

Chapter 1

Introduction

The American College of Medical Genetics Secondary Findings (ACMG-SF) list of gene-phenotype pairs is an annually updated resource of so called "actionable" genes. These genes are not related to the indication for ordering the next-generation sequencing (NGS) experiment for a patient, however, the identification of pathogenic/likely-pathogenic (P/LP) variants in said genes can be of medical value or utility to the ordering physician and the patient [1].

A recent paper established the real-world clinical relevance of screening these genes for the presence of P/LP variants in a cohort of individuals from the Icelandic population [2]. They identified that roughly 4% of the population carried an actionable genotype in the ACMG-SF gene list and that carrying such a genotype was associated with reduced lifespan [2]. The reduction in lifespan was mainly due to the carriership of P/LP variants in genes associated with cancer phenotypes. The authors concluded that screening for and relaying the actionable genotypes to individuals carrying them holds considerable potential to mitigate the disease burden on an individual level as well as on the level of society in general [2].

From a genetic perspective, the Icelandic population is specific in the sense that it is relatively small (around 300,000 individuals) and, due to the island's isolation, almost entirely descends from a single family tree [3]. Thus, the results obtained in that population are not readily translatable to other, genetically more diverse, populations.

In the current work, we therefore aimed to establish the prevalence of P/LP variants in ACMG-SF genes for the Genome Aggregation Database (gnomAD) project and compared this prevalence between the diverse set of the gnomAD populations, with the aim of establishing the genetic basis for population-based disease mitigation strategies.

Chapter 2

Theoretical background

2.1 Genetic variants

Genetic variants are permanent changes in the nucleotide sequence deviating from sequences observed in the reference genome e.g. GRCh37 [4]. The human reference genome is crucial for majority of high throughput re-sequencing-based biomedical research [5]. The assembly of a reference genome is a computational representation of a genome sequence [5, 6]. The latest human genome assembly is GRCh38, with its predecessor being GRCh37. The GRCh37 reference was extensively used in sequencing data analysis for many years and plenty of existing pipelines and annotation tools were developed based on this assembly [5].

The American College of Medical Genetics and Genomics (ACMG) recommends to classify variants using one of the following modifiers: pathogenic, likely pathogenic, uncertain significance, likely benign, or benign. While these modifiers might not encompass all human phenotypes, they provide a five-tier system to categorize variants significant to Mendelian diseases. The term 'likely' should be used where data indicates a high probability of a variant being pathogenic or benign. It is also important to mention that currently, most variants lack sufficient data to be certainly assigned to any of the four types due to the heterogeneous nature of many diseases [4].

Nonetheless, the ACMG provides criteria for classifying variants into distinct categories. Variants with very strong evidence of pathogenicity include loss-of-function (null) variants in genes where loss-of-function (LoF) is a known mechanism of disease. These variants typically can be assumed to disrupt gene function by leading to the complete absence of the gene product due to lack of transcription or nonsense-mediated decay of an altered transcript [4]. Another mechanism of disease, besides LoF, is gain-of-function (GoF). GoF variants result from genomic changes that cause the normal gene product to be inappropriately expressed, or to obtain a new abnormal function through alteration of the gene product itself [7].

It is essential to ensure that null variants align with known mechanisms of disease pathogenicity and established inheritance pattern of the disease transmission, as in some conditions null variants are not associated with clinical phenotype; for example, null variants in conditions where gain-of-function is an actual disease-causing mechanism and/or heterozygous null variants in diseases where mode of inheritance/pathogenicity requires biallelic disruption. Additionally, it is important to consider the presence of alternate gene transcripts, understand which ones are biologically relevant, and associate in which tissues the genes are expressed. I.e., it should not be assumed that a null variant will surely lead to disease if it is found in an exon where no other pathogenic variants were previously described, due to the possibility of the exon being alternatively spliced. Such considerations are crucial when predicting the pathogenicity of truncating (null) variants [4].

In addition to the properties of the variant discussed above, which in accordance to ACMG classification scheme presents 'very strong' evidence of variant pathogenicity, also other variant properties can suggest its pathogenic nature [4]. The high suggestion of pathogenicity includes in-vitro evidence of a deleterious effect of the variant on the gene or gene product; additional evidence of pathogenicity includes but is not limited to the segregation of the variant in a gene associated with the disease of interest with a phenotype highly specific for a monogenic disease. On the contrary, variants are automatically classified as benign when the allele frequency of the variant in the reference (healthy) population exceeds 5% [4].

Variants classified using the above-mentioned or alike criteria can be stored, together with classification, in different databases (i.e. ClinVar) to aid in clinical decision making [8].

2.2 ACMG recommendations for reporting of incidental findings in NGS sequencing experiments

With the increasing use of exome and genome sequencing in medicine, there exist novel opportunities to characterize diseases, individualize treatment, or conduct population screenings for disease risk. Recognizing and reporting incidental or secondary findings – results unrelated to the indication for ordering the sequencing but still potentially valuable to the physician or patient – may significantly improve patient care in these contexts. The term incidental (secondary) findings refers to unexpected positive results discovered during sequencing [9].

The American College of Medical Genetics and Genomics (ACMG) emphasizes the importance of discussing incidental findings with patients, as well as including the findings in clinical testing and reporting. The ACMG has created a list of genes and variant categories that should be reported as incidental findings. Reporting specific incidental findings can provide medical benefits for patients undergoing clinical sequencing and their families. Although all

the disorders on the list are rare, most genes and variant categories are associated with some of the more common monogenic disorders. The ACMG's recommendations are based on a consensus-driven assessment of the clinical validity and utility of reporting these findings [9].

2.3 Human populations, genetic diversity and differences in prevalence of disease

Humans are diverse genetically both among individuals and populations. In terms of populational genetics, understanding genetic variation aids in clarifying the differences and similarities between various populations, which can be defined by geographic, political, linguistic, religious, or ethnic boundaries. Worth mentioning is that only around 10% of genetic variation separates these populations [10]. Patterns of genetic diversity provide insights into population history because of the influence of various demographic events. For instance, a reduction in population size decreases its genetic diversity, while growth of population size will increase it. Additionally, the extent of genetic isolation of a population affects genetic diversity – more isolated populations will be more genetically unique [10].

Human genetic variation is shaped by the evolutionary history of the species and it shows a continuous pattern caused by demographic pressures such as mate selection, genetic drift, gene flow, and mutations. Recent genome-wide associated studies (GWAS), which can be used to localize areas of genomes that contribute to health disparities, have determined numerous genetic variants that vary significantly among human populations and are associated with disease risk. Some of these variants are specific to certain populations and may influence disease risk, helping to explain differences in disease burden between different racial or ethnic groups. By inspecting genetic ancestry, researchers are able to discover potential biological differences that could contribute to heterogeneity across racial groups [11].

Disease risk is multifactorial and affected by socio-economic, demographic, cultural, environmental, and genetic factors. Understanding global genetic diversity and its impact on health and disease can bring valuable insights into the biological mechanisms underlying disease risk. It can also aid in quantifying the impact of the interaction between genetic and environmental variations in shaping population-level disease risk. Conducting genomic research in diverse populations can improve therapeutic development and precision medicine initiatives to promote health equity [12].

There are several well-documented examples of genes increasing disease risk in specific human populations. For instance, the risk of breast cancer is greater among Jews, compared to non-Jews. Approximately 2.0% to 2.5% of Ashkenazi Jewish women carry one of three founding mutations in the BRCA1 and BRCA2 genes. Each of these mutations is associated with a

high lifetime risk of invasive breast cancer. The findings are significant for genetic screening of Jewish patients with breast cancer and for counseling their relatives [13]. Another example is the Finnish population, which has undergone several reductions in the population size, which have caused decreased genetic diversity in the population. This reduced diversity may influence the spectrum of risk variants for diseases such as breast cancer, with only a few variants potentially covering most of the pathogenic variants in the risk genes. Breast cancer is the most frequently diagnosed cancer among Finnish women [14].

Since individuals exhibit diverse and unique characteristics at the molecular, physiological, environmental, and behavioral levels, they may require personalized interventions, that are tailored to these specific characteristics. Technologies such as DNA sequencing, proteomics, or imaging protocols has aided validation of the need for tailored interventions [15].

In the following study, we will investigate the differences in the prevalence of pathogenic or likely pathogenic variants in genes from the ACMG gene list in an ethnically diverse gnomAD population, with the aim of identifying populations where tailored preventive interventions referring to diseases caused by said variants could be of public health benefit.

2.4 Related work

Venner et al. [16] examined pathogenic and likely pathogenic variants identified in the All of Us cohort. The All of Us Research Program from the National Institutes of Health (NIH) is collecting data including whole-genome sequences, health records, and surveys from a million participants with diverse ancestry and access to healthcare. Venner et al. emphasized the lack of diversity in large genomics cohorts, noting that most sequencing studies have focused on populations of European ancestry. This is problematic, due to much of pathogenic variation being specific to ancestral populations. To address this issue, a comprehensive collection of genomic data from diverse populations is necessary [16].

In their study, they found that the European ancestry subgroup had the highest overall rate of pathogenic variation, with 2.26% of participants possessing a pathogenic variant. Other ancestry groups had lower rates of pathogenic variation, including 1.62% for the African ancestry group and 1.32% in the Latino/Admixed American ancestry group [16]. Venner et al. also compared the results (rates of pathogenic findings) from the All of Us cohort with gnomAD [16].

For their research, Venner et al. used the VIP database to annotate known pathogenic or likely pathogenic (P/LP) variants. This database is a collection of variants identified during clinical reporting at the Human Genome Sequencing Center-Clinical Laboratory. The variant data of the All of Us cohort was generated using the GRCh38 human reference build. They focused on

coding regions for the 73 genes listed in the ACMG v3.0 list by subsetting the whole genome variants. Additionally, they annotated aggregate variant data using the Variant Effect Predictor and used ClinVar annotation for the gnomAD database to compare these results with the All of Us cohort [16].

Jensson et al. [2] evaluated the prevalence of coding and splice variants in genes on the ACMG Secondary Findings list within the genomes of Icelanders. They determined the pathogenicity of all reviewed variants using evidence reported in the ClinVar database, the frequency of variants, and their associations with disease, resulting in a manually curated set of actionable genotypes (variants). The study found that individuals carrying actionable genotypes had shorter shorter median survival compared to noncarriers. Particularly, carrying an actionable genotype in a cancer gene was associated with survival reduced by 3 years compared to noncarriers [2].

In their research, Jensson et al. aligned reads to the human genome assembly GRCh38 and used the Variant Effect Predictor for annotation of variants. To assess pathogenicity, they cross-referenced the variants with the ClinVar database [2].

Chapter 3

Methods

3.1 Population, bioinformatic tools and databases

3.1.1 gnomAD

The Genome Aggregation Database (gnomAD) aggregates exome and genome sequencing data from various large-scale sequencing projects, making the data publicly available. The sequences come from individuals across different populations worldwide [17].

The gnomAD database is periodically updated, and in the following study, we used version 2.1. We downloaded the file of all chromosome sites in Variant Call Format (VCF) from the gnomAD website [18]. VCF is a format able to store SNVs, insertions, deletions, and structural variants along with their incomplete annotations. It can represent diverse genomic variations related to a single reference sequence and is designed to handle extensive genotype data and annotations [19].

This version of the gnomAD database includes data from 125,748 exomes and 15,708 genomes from human sequencing studies [17]. The database categorizes data into seven populations, with two further divided into subpopulations: African/African American, Latino/Admixed American, Ashkenazi Jews, East Asian (subdivided into Koreans and Japanese), Finnish, Non-Finnish European (subdivided into Bulgarian, Estonian, North-Western European, Southern European and Swedish) and South Asian. Our study focused on genomic data, hence populations with exome-only information were excluded – because of that we did not incorporate data from Korean, Japanese, Bulgarian, Swedish, and South Asian populations. The populations used and the respective numbers of genomes included from each are shown in Table 3.1.

Table 3.1: The table below presents the number of genomes included in the gnomAD database per each population, along with corresponding numbers of alleles per population. The populations are, in order: Latino/Admixed American (amr), Ashkenazi Jews (asj), African/African American (afr), East Asian (eas), Finnish (fin), North-Western European (nwe), Southern European (seu), Estonian (est) and Other Non-Finnish European (onf). Additionally, it shows the medians of total number of alleles in samples for each ACMG-listed gene associated with cancer phenotypes, per population. As each individual carries two alleles of each gene, due to humans being diploid organisms, this results in two alleles of each gene per individual.
Moreover, the medians vary slightly, possibly because of the incompleteness of sequenced data or the low quality of some alleles that lead to their exclusion. However, the values still oscillate around the expected number of alleles as the overall sequencing coverage and quality are consistent across the database.

Population	**amr**	**asj**	**afr**	**eas**	**fin**	**nwe**	**seu**	**est**	**onf**	**Total**
Nr of genomes	424	145	4359	780	1738	4299	53	2297	1069	15164
Nr of alleles	848	290	8718	1560	3476	8598	106	4594	2138	30328
Median for APC	848	290	8708	1560	3472	8594	106	4582	2136	30296
Median for BMPR1A	848	290	8714	1558	3476	8596	106	4590	2138	30316
Median for BRCA1	848	290	8714	1558	3474	8596	106	4588	2137	30311
Median for BRCA2	848	290	8712	1560	3474	8592	106	4588	2136	30306
Median for MAX	845	290	8700	1558	3466	8594	106	4556	2136	30251
Median for MEN1	848	290	8718	1558	3474	8596	106	4574	2138	30302
Median for MLH1	845	288	8693	1556	3465	8590	106	4546	2134	30223
Median for MSH2	848	290	8704	1558	3472	8594	106	4586	2138	30296
Median for MSH6	848	290	8696	1554	3452	8588	106	4550	2130	30211
Median for MUTYH	848	290	8710	1558	3474	8596	106	4586	2138	30306
Median for NF2	848	290	8710	1560	3472	8596	106	4586	2138	30306
Median for PALB2	848	290	8712	1560	3474	8596	106	4590	2138	30314
Median for PMS2	845	290	8702	1554	3464	8589	106	4575	2134	30259
Median for PTEN	840	290	8690	1558	3468	8528	106	4554	2130	30164
Median for RB1	834	288	8592	1552	3218	8528	106	4282	2104	29504
Median for RET	848	290	8697	1557	3468	8584	106	4555	2136	30241
Median for SDHAF2	848	290	8704	1558	3474	8598	106	4588	2138	30308
Median for SDHB	848	290	8708	1560	3474	8596	106	4590	2136	30310
Median for SDHC	826	284	8590	1552	3102	8486	102	4466	2100	29508
Median for SDHD	832	290	8594	1544	3396	8562	104	4454	2118	29894
Median for SMAD4	846	290	8708	1558	3444	8584	106	4564	2134	30234
Median for STK11	848	290	8690	1560	3458	8590	106	4550	2132	30224
Median for TMEM127	848	290	8712	1560	3476	8596	106	4590	2138	30316
Median for TP53	-	-	-	-	-	-	-	-	-	-
Median for TSC1	848	290	8709	1559	3472	8594	106	4583	2136	30297
Median for TSC2	848	290	8704	1560	3470	8594	106	4578	2138	30294
Median for VHL	846	290	8704	1560	3466	8596	104	4576	2136	30278
Median for WT1	848	290	8714	1560	3476	8596	106	4586	2138	30314

3.1.2 BCFtools

BCFtools is a component of the SAMtools package, designed for handling Variant Call Format (VCF) and Binary Call Format (BCF) files. BCFtools component infers various statistics from these files. For efficiency, this tool can be used on files in BCF format, which is the binary representation of VCF, making it more compact and faster to process than VCF. In addition to manipulation with VCF/BCF data, this tool can compute various statistics directly from sequencing data [20].

BCFtools includes various commands, such as SNP/indel calling, file concatenation, fixed-threshold filtering, file transformation into user-defined formats, or file subsetting and conversion [21]. Most commands allow filtering sites either by a region, list of sites, or a general Boolean expression involving VCF tags (–include, –exclude options). The 'bcftools view' command provides conversion between the text VCF and the BCF formats, with both formats supporting either plain (uncompressed) or block-compressed (BGZF) versions for random access and compact size. Arbitrary fields can be extracted and formatted into a custom text output using a query feature, which is particularly useful for scripting. BCFtools programs are written in the C programming language and optimized for low memory consumption and high speed [22]. In this project, we mostly used BCFtools (version 1.15.1) for manipulation of the gnomAD VCF file.

3.1.3 VEP

The Ensembl Variant Effect Predictor (VEP) is an open-source and freely accessible tool designed to analyze and annotate genomic variants in both coding and non-coding regions of the genome. It can be used in various types of research or clinical contexts, aiding in the process of interpreting different types of genetic variants [23]. In the following work, we used release 110 of VEP.

3.1.4 ClinVar

ClinVar is an open and free public repository that provides detailed information about connections between variants in the human genome and the phenotypes (diseases) associated with those variants, supported by scientific evidence. It accepts detailed submissions from both individual users and organizations, which include information about the phenotype, functional significance, clinical significance, and methods used for identifying variants. Submissions are classified based on criteria, such as the method of data collection and the review level. The submissions encompass variations discovered through clinical testing, research, and curated literature. Genomic variation plays a crucial role in ClinVar's data structure, particularly in

showing the connection between variations and phenotypes. ClinVar can represent the interpretation of a single allele, compound heterozygotes, haplotypes, and combinations of alleles across various genes [24].

In this research, we used the tab-delimited variant summary file (version from November 2, 2023), available on the ClinVar FTP site (link to the site [25]). This report describes each variant at a specific location on the genome for which data have been submitted to ClinVar. It contains information such as clinical significance, a list of condition names, the number of submitters describing a given variant, or review status for the aggregate germline classification [26]. We downloaded this file to annotate the gnomAD database, to further incorporate information regarding the clinical significance of variants.

3.1.5 ACMG gene list

The ACMG gene list provides recommendations for a minimum set of gene-phenotype pairs for opportunistic screening to identify and manage risks for selected genetic disorders through established interventions that prevent or significantly reduce morbidity and mortality. The genes on the list are categorized into categories based on their associated phenotype: cancer, cardiovascular, inborn errors of metabolism and miscellaneous [27]. The latest version, used in this research project, is 3.2, containing 81 genes. A link to the NCBI website with the adapted full table is available at [28].

3.2 Bioinformatics pipeline

The bioinformatics pipeline presented in Figure 3.1, begins with the extraction of ACMG-listed gene positions from the input VCF file (the gnomAD database). Then, the file underwent annotation using VEP and ClinVar for additional information. Lastly, the file was filtered according to several criteria, resulting in two files that later were statistically analyzed. The pipeline is presented in detail further in this following section 3.2. The methods were inspired by Venner et al. [16] and Jensson et al. [2], with some alterations. Similarly to these projects, we used the VEP and ClinVar for variant annotation and focused on the regions of genes listed in the ACMG list.

The implementation of the methods described in this section can be found at:
`https://github.com/aleksandrabaum/thesis_project`

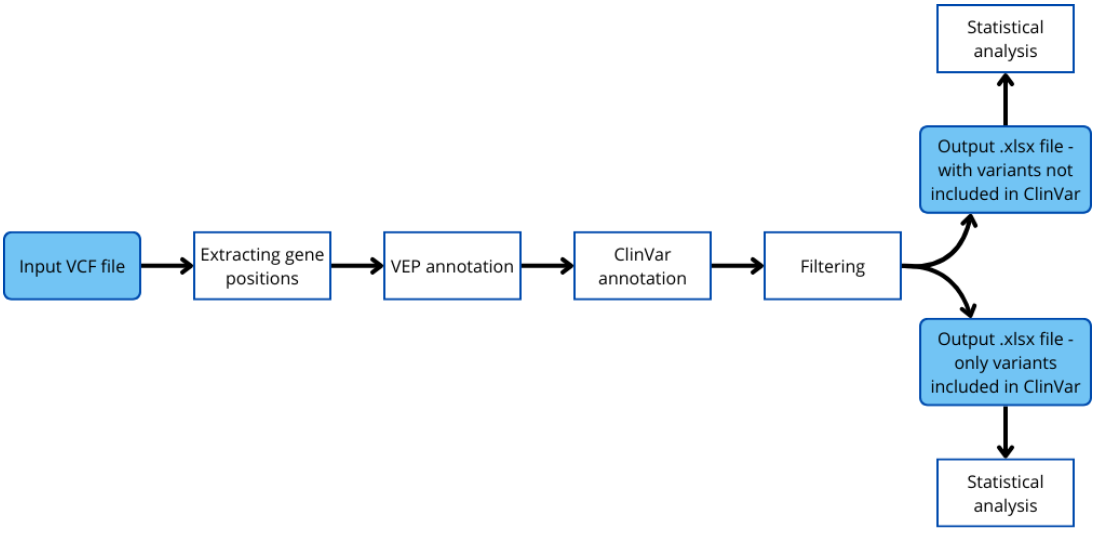

Figure 3.1: Diagram showing the full bioinformatics pipeline.

To incorporate the methods presented below, we had to use coding to some extent. Mainly, coding bits were executed in a Jupyter Notebook using Python programming language (version 3.10).

In addition to Python, some commands were executed using bash commands, which aid quicker manipulation of big files. Bash is the shell and command language interpreter for the GNU operating system. In interactive mode, shells accommodate input from the keyboard, while in non-interactive mode, they execute commands from a file [29]. In this project, we used simple Bash commands, which are highlighted in this section.

3.2.1 Extracting gene positions and filtering

The genomic database obtained from the gnomAD website contained extensive information on human variants. However, the size of the database, with over 460 GB of data, made it impractical to directly manipulate it. Thus, it required filtering and size reduction in order to smoothly handle the data.

Firstly, a file containing the regions of interest, so regions corresponding to genes listed in the ACMG secondary findings list, was created. It was done by retrieving the list of genes recommended by ACMG from the NCBI website's table (link to the website [28]). In the script, we scraped the Ensembl website to obtain the chromosome, start, and stop position for each gene which was then formatted into a .bed file. The BED format is used to store genomic coordinates and optionally their associated annotations if needed. The coordinates consist of the chromosome, start position and end position, separated by either tabs or spaces [30]. In this project, we used the GRCh37 version of the genome assembly and based all

genomic positions accordingly. To ensure comprehensive coverage of variant positions of the genes listed by ACMG, we used positions 10,000 nucleotides before and after each gene's start and stop positions.

The .bed file was used as a reference to extract only variants within the specified regions from the 460 GB gnomAD genomic database. This process was done by utilizing the 'bcftools view' command with '--regions_file' flag and specifying the aforementioned file containing regions of interest. As a result, a considerably smaller file was generated, with 49 GB of data. Furthermore, the file was indexed with 'bcftools index' command to create a .csi (coordinate-sorted index) file to enable random access – it is required by bcftools for further usage of a compressed .vcf file.

A considerable amount of data in the already reduced database could be further simplified for easier manipulation, particularly within the *INFO* column of the .vcf database, including diverse information about populations in the gnomAD database. To address this, the file was transformed into a different format using 'bcftools query' command. The resulting file comprised of tab-delimited columns: *CHROM*, *POS, ID, REF, ALT, QUAL, FILTER* and *INFO/AF* (which included allele frequency of the variant). This format included less data per variant, which led to reducing the size of the file to just under 30 GB.

3.2.2 Annotating the file using VEP

In the next step, the file was annotated using the VEP tool. While basic annotation had been already conducted for the variants in the gnomAD database, this additional annotation added information crucial for filtering out unwanted variants.

Annotation of such a big file was the most consuming step of preparation, eventually resulting in a 35 GB file containing over 34 million variants. The next filtering steps ideally could have been conducted effortlessly using the pandas dataframe, however, the annotated, reduced gnomAD database was still too large for direct loading into a dataframe. Therefore, the file was divided into four smaller files using the bash command 'split -l 9000000 {file name}' directly in the terminal, ensuring each file had approximately 9 million lines.

In the next phase, each of the smaller files could easily be uploaded into pandas dataframes. One by one, the files were loaded into pandas, with variants exceeding 5% allele frequency being filtered out as these are considered benign according to ACMG criteria. Additionally, in order to achieve this, the allele frequency column's data type was changed from string values to numeric values. The files remained as separate and were saved in tab-delimited formatting in .csv files.

3.2.3 Annotating the file with ClinVar database

Using the extensive information from the ClinVar database, each of the four files underwent a left join operation. The ClinVar variant summary database was imported into a dataframe, and variants of the GRCh38 assembly were left out. Left-joining this database aided filtering, leaving only variants categorized as having a 'HIGH' impact according to VEP or having the clinical significance of 'Pathogenic', 'Likely pathogenic', or 'Pathogenic/likely pathogenic' according to ClinVar. This significantly reduced the size of each file, allowing them to be combined into one .csv file via the bash command 'cat {all 4 file names} > combined.csv' in the terminal, eventually saved as an Excel spreadsheet (.xlsx).

Further, using a custom function, information from the *INFO* column was parsed into separate columns within the dataframe. This improved clarity and enabled easier filtering or information retrieval. With that it was also possible to filter out the NMD-escaping variants from variants with 'HIGH' impact. Nonsense-mediated mRNA decay (NMD) is a translation-dependent mRNA surveillance mechanism in eukaryotes that preserves the quality of gene expression. In humans, NMD prevents the synthesis of potentially harmful proteins from mutated mRNAs [31].

The file underwent additional filtering. For variants present in the ClinVar database, we kept only those submitted at least twice using the *NumberSubmitters* column. We excluded variants with 'Conflicting interpretations of pathogenicity' in the *ClinicalSignificance* column and kept variants classified as 'Pathogenic', 'Pathogenic/Likely pathogenic', 'Likely pathogenic', 'Pathogenic; drug response' or 'Likely pathogenic; drug response'. We also included high-impact variants not reported in ClinVar. This filtering resulted in one of the output .xlsx files used for statistical analysis of variants associated with cancer phenotypes. Then, to analyze all four phenotypes, we included only variants with the aforementioned clinical significance values reported in the ClinVar database and saved them to a separate .xlsx file. Both .xlsx files were then analyzed as described in section 3.3.

3.3 Statistical analysis

Extracting information regarding the populations included in gnomAD was essential for clarity and further computations. The most crucial were details about the number of individuals tested for a given variant within each population and how many of them actually had the causative variant in a gene of interest. These metrics were stored in the *INFO* column of gnomAD database as AN (representing 'Total number of alleles in samples') and AC (representing 'Alternate allele count for samples' – causative variant) respectively. For each of the nine populations of interest, these values were segregated into distinct columns within the dataframe.

Then, the AC/AN ratio for each variant in the dataframe was computed to determine the frequency of the variant within each population. The frequency values were included in separate columns for each population.

Further, to understand the frequency of pathogenic variants in genes across populations, variant information had to be grouped by genes, calculating the frequency of all pathogenic variants occurring in each gene of interest in a given population. Variants were grouped using the *SYMBOL* column, which contains gene names. For each gene group, the median of AN values of variants, along with the sum of AC values of variants, were calculated. Again, the frequency of each gene's appearance was computed by dividing the sum of AC values by the median of AN values (in a per gene manner). The resulting values were stored in a dataframe, one for each population, with the following columns: *SYMBOL*, *AC_{population name}*, *AN_{population name}* and *AC/AN_ratio_{population name}*. These dataframes were saved as a dictionary, with population names as keys and their respective dataframes as values.

Subsequently, data was prepared for Fisher's exact test, which determines whether the proportions of data described by two or more categorical variables are random [32]. We needed the sum of AC values per gene and the difference between the sum of AC values and the median of AN values per gene. This information was stored in a dictionary, where population names were keys and the corresponding values were lists containing gene names, sums of AC values of variants of the gene, medians of AN values of variants of the gene, and the differences between these two values.

During this process, it was observed that while creating the reduced gnomAD database, in which only the regions of ACMG-reported genes plus 10,000 nucleotides on either side of those regions were kept, it resulted in the addition of variants related to genes beyond the ACMG secondary findings list. Some of these variants were still present in the annotated and filtered file. To address this, the previously web-scrapped list of ACMG secondary findings from the NCBI website was utilized to filter each dataframe for each population, leaving only entries with gene names from the ACMG list in the *SYMBOL* column. This process could not be done at the beginning due to the impracticality of filtering the entire gnomAD database by gene names. Instead, it was more practical to create a regions file with the regions of interest and use that for database filtering.

As previously mentioned, Fisher's exact test was performed only for genes listed in the ACMG secondary findings. Not all of the genes listed by ACMG had corresponding variants in the annotated and filtered database, thus any gene absent from the dictionary prepared for the Fisher's exact test was omitted. The test was conducted for each of the 45 possible population pairs. For each pair, a contingency table was created using the sum of AC values for populations A and B, along with the differences between AN and AC values for these populations. Based

on this, the odds ratio and p-value were calculated. Results were stored in a nested dictionary, with gene names as keys, and for each gene, another dictionary including population pairs and their corresponding Fisher's exact test results.

Next, Fisher's exact test was computed for groups of genes based on their associated phenotypes. The phenotypes with the corresponding genes were drawn from Miller et al. [27]: genes related to cancer phenotypes, cardiovascular phenotypes, inborn errors of metabolism phenotypes, and miscellaneous phenotypes. The phenotypes and their respective genes were written into a .txt file, which was the basis for a dictionary with phenotypes as keys and lists of associated genes as values. Another nested dictionary was created, with populations as keys and, for each phenotype group, the sum of AC values for all genes belonging to the given group and the difference between the median of AN values and the sum of AC values for the group.

The Fisher's exact test was then performed for the phenotype-based gene groups. Similar to the gene-specific test, a contingency table was created for each population pair, calculating the odds ratios and p-values.

To visualize the results, frequencies of variants were computed within each gene within each phenotype group, using the Python visualization library matplotlib. For the gene-specific calculations, variants were grouped by gene, summing AC values and calculating medians of AN values, and then dividing these values to find frequencies. These frequencies were plotted with populations on the x-axis and frequencies on the y-axis. The same process was applied to phenotype groups. Plots for phenotype groups were generated in the same manner, with populations on the x-axis and frequencies on the y-axis.

When visualizing pairwise comparisons, only the comparisons between populational frequencies (gene/phenotype-wise) that had a p-value of <0.05 were visualized (this value was also considered statistically significant).

Chapter 4

Results

The pipeline resulted in notable reduction in the number of variants for statistical analysis in comparison to the gnomAD database which includes only variants in regions of ACMG-listed genes. Ultimately, we were left with 12,358 variants of pathogenic/likely pathogenic significance or high impact, associated with 69 ACMG-listed genes. After excluding variants with fewer than 2 submitters, there retained 11,426 variants, still associated with 69 ACMG-listed genes. Further filtering excluded the high-impact variants classified in ClinVar as benign, likely benign, or of uncertain significance. This step left us with 11,337 variants from 63 ACMG-listed genes. Finally, excluding variants not reviewed by ClinVar reduced the sample to 1338 variants from 48 ACMG-listed genes.

The statistical analysis of the results is described in more detail in the following chapter, accompanied by relevant plots and tables. The chapter is organized into sections for each of the analyzed phenotypes: cancer phenotypes, cardiovascular phenotypes, inborn errors of metabolism phenotypes, and miscellaneous phenotypes. Most genes associated with cancer phenotypes follow a loss-of-function pathogenic mechanism of disease [33, 34]. Therefore, for cancer phenotypes, the experiment was broken down into two parts: first (CV_P+HIGH_VEP), analyzing variants reviewed in ClinVar as pathogenic or likely pathogenic and variants with high impact, excluding high-impact variants that were reviewed as benign, likely benign or of uncertain significance or with conflicting interpretations, and second (CV_P), analyzing only variants reported in ClinVar as pathogenic or likely pathogenic, excluding variants not present in this database. For all other phenotypes, since gain-of-function mechanisms of disease can also be a significant pathogenic mechanism contributor, we used information from the CV_P experiment, utilizing only data already available in ClinVar that was reviewed as pathogenic or likely pathogenic.

In this chapter, we present the frequencies of individuals carrying pathogenic/likely pathogenic variants in genes associated with each of the four phenotypes to establish the susceptibility to

various genetic variations in different populations. We also identify the genes and their variants that drive the high frequencies. The variants are presented with position IDs in the format: chromosome-position-reference_allele-alternate_allele.

The resulting plots described in this section, along with additional plots showing frequencies of pathogenic/likely pathogenic variants per each gene, can be found at:
`https://github.com/aleksandrabaum/thesis_project/tree/main/figures`

4.1 Cancer phenotypes

In this section, there are presented the results of the analysis of variants from genes associated with cancer phenotypes. The respective genes and frequencies of pathogenic/likely pathogenic variants are shown in Table 4.1. The table highlights which populations have higher or lower frequencies of certain genetic variants and outlines differences between the two experiments performed for variants associated with cancer phenotypes.

Table 4.1: This table presents genes associated with cancer phenotypes, describing their modes of inheritance, mechanisms of disease and frequencies of pathogenic/likely pathogenic or high-impact variants by each population. The populations are namely: Latino/Admixed American (amr), Ashkenazi Jews (asj), African/African American (afr), East Asian (eas), Finnish (fin), North-Western European (nwe), Southern European (seu), Estonian (est) and Other Non-Finnish European (onf). In the mode of inheritance column, AD refers to autosomal dominant and AR to autosomal recessive. In the mechanism of disease column, LoF represents the loss-of-function mechanism and GoF represents the gain-of-function mechanism. The frequencies of pathogenic variants include information from both analyses conducted for cancer phenotypes. First value was computed using variants reviewed in ClinVar as pathogenic or likely pathogenic, excluding high-impact variants reviewed as benign, likely benign or of uncertain significance (while keeping the high-impact variants not reviewed by ClinVar). The second value was computed using only variants present in ClinVar database and reported as pathogenic or likely pathogenic. The mode of inheritance was incorporated from Miller et al. [27], while the mechanisms of disease were drawn by searching through GeneReviews [33] and OMIM databases [34]. GeneReviews offers clinically relevant and medically actionable information for inherited conditions in a standardized journal-style format, covering aspects such as diagnosis, management, and genetic counseling for patients and their families [33]. The Online Mendelian Inheritance in Man (OMIM) is a comprehensive and authoritative resource on human genes and genetic phenotypes, containing information on all known mendelian disorders and over 16,000 genes [34].

Disease Group and Gene	Mode of inheritance	Mechanism of disease	Frequency of pathogenic variants								
			amr	asj	afr	eas	fin	nwe	seu	est	onf
Cancer phenotypes											
APC	AD	LoF	0.0/-	0.0/-	0.0/-	0.0/-	0.0/-	0.0/-	0.0/-	0.000218/-	0.000936/-
BMPR1A	AD	Unknown	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-
BRCA1	AD	LoF	0.0/0.0	0.0/0.0	0.000574/0.000574	0.0/0.0	0.0/0.0	0.000582/0.000582	0.009434/0.009434	0.003269/0.003269	0.000935/0.000935
BRCA2	AD	LoF	0.001179/0.001179	0.006897/0.006897	0.002296/0.002296	0.001923/0.001923	0.0/0.0	0.001629/0.001629	0.0/0.0	0.00109/0.00109	0.000936/0.000936
MAX	AD	LoF	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-
MEN1	AD	LoF	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-
MLH1	AD	LoF	0.001773/-	0.0/-	0.0/-	0.0/-	0.0/-	0.0/-	0.0/-	0.0/-	0.0/-
MSH2	AD	LoF	0.0/0.0	0.003448/0.003448	0.0/0.0	0.0/0.0	0.0/0.0	0.000116/0.0	0.0/0.0	0.000218/0.000218	0.0/0.0
MSH6	AD	LoF	0.0/0.0	0.0/0.0	0.002185/0.000346	0.0/0.0	0.001153/0.0	0.001164/0.000233	0.0/0.0	0.003516/0.0	0.001406/0.0
MUTYH	AR	LoF	0.0/0.0	0.0/0.0	0.000115/0.000115	0.0/0.0	0.0/0.0	0.000465/0.000465	0.0/0.0	0.000218/0.000218	0.000935/0.000935
NF2	AD	LoF	0.0/-	0.0/-	0.0/-	0.0/-	0.000288/-	0.0/-	0.0/-	0.0/-	0.0/-
PALB2	AD	LoF	0.0/0.0	0.0/0.0	0.00023/0.00023	0.000641/0.000641	0.000864/0.000864	0.0/0.0	0.0/0.0	0.0/0.0	0.000468/0.000468
PMS2	AD	LoF	0.0/0.0	0.0/0.0	0.000345/0.000345	0.000643/0.000643	0.0/0.0	0.000931/0.000931	0.0/0.0	0.000654/0.000654	0.001404/0.001404
PTEN	AD	LoF	0.0/0.0	0.0/0.0	0.000115/0.000115	0.0/0.0	0.0/0.0	0.000117/0.000117	0.009434/0.009434	0.0/0.0	0.0/0.0
RB1	AD	LoF	0.0/0.0	0.0/0.0	0.000121/0.0	0.0/0.0	0.0004/0.0	0.000367/0.000117	0.0/0.0	0.001872/0.0	0.0
RET	AD	GoF	0.001179/0.001179	0.0/0.0	0.000115/0.000115	0.0/0.0	0.0/0.0	0.000349/0.000349	0.0/0.0	0.000439/0.000439	0.000468/0.000468
SDHAF2	AD	LoF	0.0/-	0.0/-	0.000115/-	0.0/-	0.0/-	0.0/-	0.0/-	0.0/-	0.0/-
SDHB	AD	LoF	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-
SDHC	AD	LoF	0.0/0.0	0.0/0.0	0.000115/0.000115	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
SDHD	AD	LoF	0.004894/-	0.008197/-	0.00471/-	0.000728/-	0.01198/-	0.003258/-	0.0/-	0.006964/-	0.001172/-
SMAD4	AD	Unknown	0.001312/-	0.0/-	0.000354/-	0.0/-	0.004003/-	0.000712/-	0.0/-	0.00299/-	0.000479/-
STK11	AD	LoF	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-
TMEM127	AD	LoF	0.0/0.0	0.0/0.0	0.00023/0.00023	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.000218/0.0	0.0/0.0
TP53	AD	LoF	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-
TSC1	AD	LoF	0.0/0.0	0.0/0.0	0.000115/0.000115	0.0/0.0	0.0/0.0	0.000116/0.0	0.0/0.0	0.0/0.0	0.0/0.0
TSC2	AD	LoF	0.001179/0.001179	0.0/0.0	0.000115/0.000115	0.0/0.0	0.0/0.0	0.000465/0.000465	0.0/0.0	0.0/0.0	0.0/0.0
VHL	AD	LoF	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.000116/0.000116	0.0/0.0	0.0/0.0	0.0/0.0
WT1	AD	LoF	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-

4.1.1 Variants reviewed as pathogenic/likely pathogenic or with high impact

Figure 4.1 shows the plot of frequencies of individuals carrying pathogenic/likely pathogenic variants or high-impact variants in genes associated with cancer phenotypes across various populations. Southern European and Estonian populations exhibit the highest frequency of pathogenic/likely pathogenic variants or high-impact variants, reaching a frequency of almost 2%. Although the frequency of these variants for the Southern European population is high, these differences are not statistically significant compared to proportions in other populations, due to the small size of the Southern European population sample. For genes associated with cancer phenotypes, this frequency for the Estonian population diverges significantly from several other populations, namely African/African American, East Asian, Finnish, North-Western European, and Other Non-Finnish European. Low p-values indicate the statistical significance of the differences in frequencies of pathogenic/likely pathogenic or high-impact variants. A highly significant difference with a p-value below 0.0001 was observed for Estonian and North-Western European populations. Another population reaching a high frequency of pathogenic/likely pathogenic variants in genes associated with cancer phenotypes is Ashkenazi Jews with around 1.7% frequency. The East Asian population reaches the lowest frequency out of the examined populations.

For the three populations with the highest frequencies of pathogenic/likely pathogenic variants, we identified the gene and variant driving the high frequency. In all three populations, Estonian, Ashkenazi Jews, and Finnish, it was the same gene, SDHD, and the same variant, with position ID 11-111990013-A-T. The variant, although having a high impact, is not reported in ClinVar database. The frequency of this variant, as well as of other variants driving high frequencies, is stated in Table 4.4.

The SDHD (succinate dehydrogenase complex subunit D) gene encodes a component of complex II in the respiratory chain. Mutations in this gene are linked to tumor formation, including hereditary paraganglioma. Individuals with pathogenic variants in SDHD have an increased risk of developing paraganglioma-pheochromocytoma tumors, especially in the head and neck region. More data is needed to assess tumor risk for individuals with SDHD pathogenic variants and optimize surveillance guidelines [35].

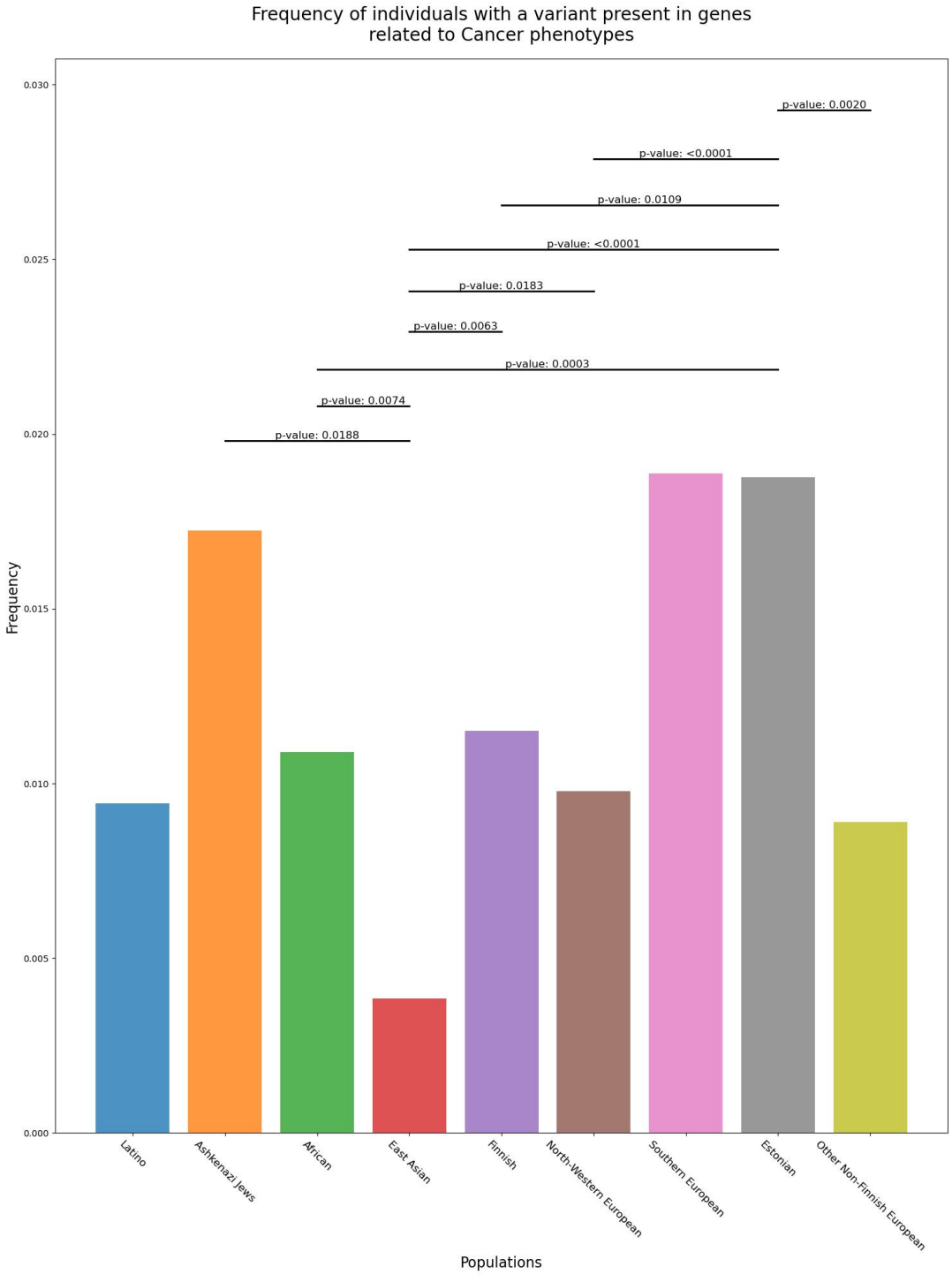

Figure 4.1: Bar plot showing the frequency of individuals carrying pathogenic/likely pathogenic or high-impact variants in genes associated with cancer phenotypes across different populations. These results are derived from the filtering step that excluded any variants classified as benign, likely benign, or of uncertain significance. The plot includes p-values that indicate the statistical significance of the differences in frequencies, derived from Fisher's Exact Test.

4.1.2 Only variants reviewed as pathogenic/likely pathogenic

Figure 4.2 shows the frequencies of individuals carrying only variants reviewed by ClinVar as pathogenic or likely pathogenic. This excludes variants reported as high impact, but not reported in the ClinVar database, or variants with high impact but reported in ClinVar as benign/likely benign or of uncertain significance. This approach guarantees that the only preserved variants were the ones with high confidence of association with disease. The Southern European population again reaches the highest frequency, however, it is not significantly different in frequency due to the small size of the sample. Another population with a high frequency of pathogenic/likely pathogenic variants is Ashkenazi Jews, with a frequency exceeding 1%. The Finnish population exhibits the lowest frequency of pathogenic/likely pathogenic variants, resulting in significant divergence from Ashkenazi Jews, African/African American, North-Western European, Southern European, Estonian (the lowest p-value), and Other Non-Finnish European.

For the three populations with the highest frequencies of pathogenic/likely pathogenic variants, we again identified the gene and variant driving the high frequency. In the Ashkenazi Jewish population, the gene driving a high frequency is BRCA2 and its variant with position ID 13-32914437-GT-G. In the Estonian population, the gene is BRCA1, with the variant 17-41243512-CT-C. For the Other Non-Finnish European population, the gene is PMS2, with the variant 7-6022516-C-T. We also analyzed the Finnish population to compare it with the previous result that included the variants not reported in ClinVar. In this population, the gene driving the high frequency is PALB2, with its variant 16-23646274-CA-C. The frequencies of these variants are stated in Table 4.4.

Inherited mutations in BRCA1 (BRCA1 DNA repair associated) and BRCA2 (BRCA2 DNA repair associated), significantly increase the lifetime risk of developing breast or ovarian cancer. Both BRCA1 and BRCA2 are involved in the maintenance of genome stability. Both BRCA1 and BRCA2 are considered tumor suppressor genes [36, 37]. Mutations in BRCA1 are responsible for approximately 40% of inherited breast cancers and more than 80% of inherited breast and ovarian cancers [36]. PMS2 (PMS1 homolog 2, mismatch repair system component) encodes a protein that is a key component of the mismatch repair system, correcting DNA mismatches and small insertions and deletions that can occur during DNA replication and homologous recombination. Mutations in this gene are associated with hereditary nonpolyposis colorectal cancer (Lynch syndrome) [38]. PALB2 (partner and localizer of BRCA2) encodes a protein that may function in tumor suppression by binding to the breast cancer 2 early onset protein (BRCA2) and likely permitting the stable intranuclear localization and accumulation of BRCA2 [39].

In comparison to the results presented in Figure 4.1, the frequencies for all populations dropped, with some having much more significant decreases than others. The Ashkenazi Jewish popula-

tion retained a relatively high frequency, compared to other populations, in both experiments, with a decrease in the frequency from around 1.7% to just above 1%. All other populations saw a drop of around 0.5%. Additionally, the East Asian population retained its low frequency throughout both experiments, with a difference of less than 0.1%. However, the Estonian population experienced a significant decrease of the frequency, from almost 2% to just above 0.5%.

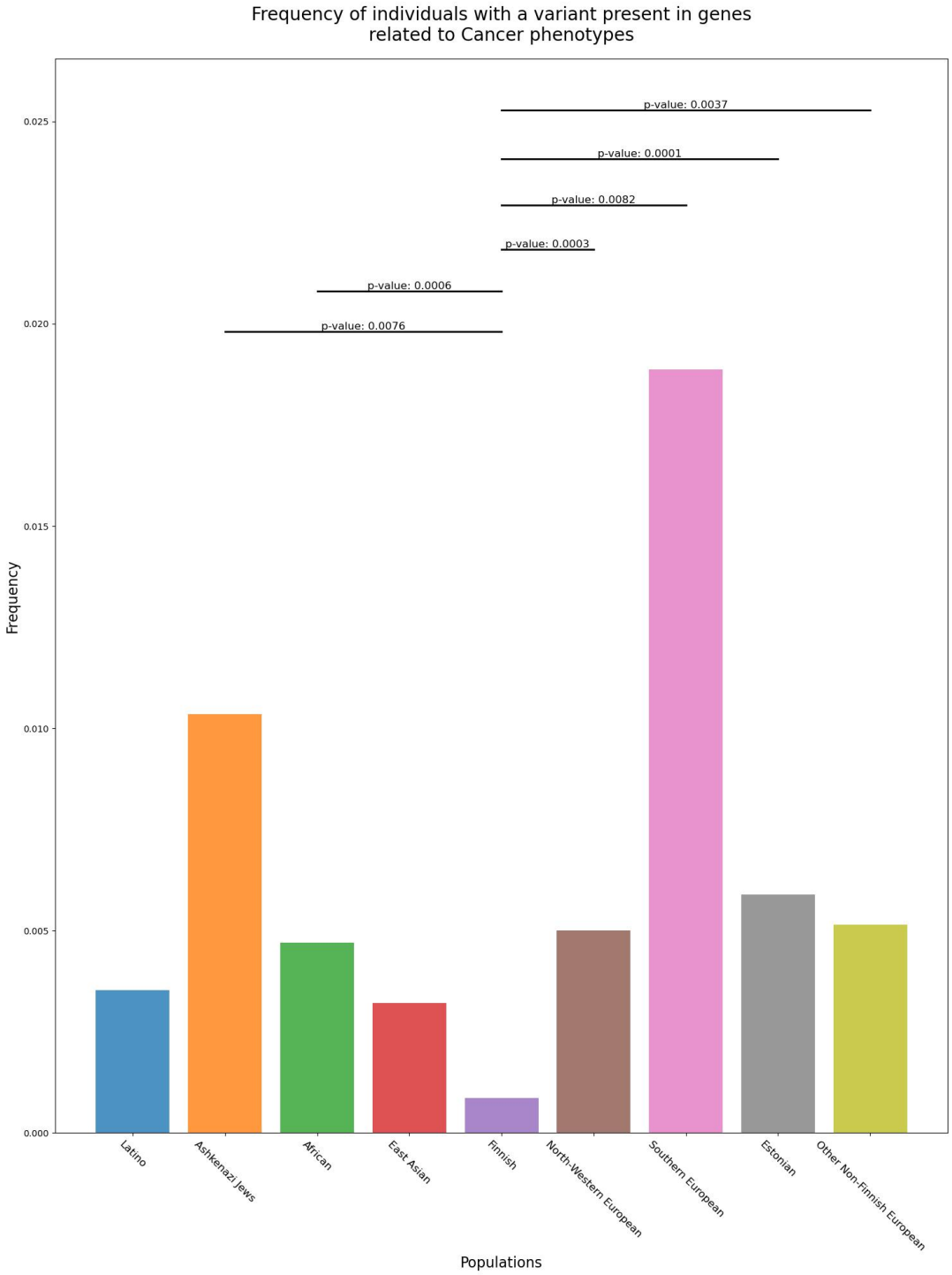

Figure 4.2: Plot showing the frequency of individuals carrying pathogenic/likely pathogenic variants in genes associated with cancer phenotypes across different populations. These results are derived from the filtering step that excluded any variants classified as benign, likely benign, or of uncertain significance and variants not reported in the ClinVar database. The plot includes p-values that indicate the statistical significance of the differences in frequencies, derived from Fisher's Exact Test.

4.2 Cardiovascular phenotypes

This section presents the results of the analysis of variants from genes associated with cardiovascular phenotypes. Table 4.2 and Table 4.3 provide additional information regarding these genes. Due to the pathogenic mechanisms of diseases of cardiovascular phenotypes, we included only the variants reported in ClinVar as pathogenic/likely pathogenic.

Figure 4.3 shows the frequencies of individuals carrying pathogenic/likely pathogenic variants in genes associated with cardiovascular phenotypes across various populations. The Other Non-Finnish European population exhibits the highest frequency of pathogenic/likely pathogenic variants, with a value exceeding 1.25%. This frequency is significantly different from frequencies observed in the Latino/Admixed American, African/African American, East Asian, Finnish and Estonian populations. Highly significant differences, with p-values below 0.0001, can be observed between Estonian and Other Non-Finnish populations as well as between North-Western European and Estonian populations. The lowest frequency of pathogenic/likely pathogenic variants is exhibited by the Estonian population.

For the three populations with the highest frequencies of pathogenic/likely pathogenic variants, we again identified the gene and variant driving the high frequency. In the Other Non-Finnish European population, the gene driving a high frequency is LDLR and its variant with position ID 19-11227604-G-A. In the North-Western European population, the gene is also LDLR, but two different variants, 19-11231156-G-A and 19-11216133-G-A, contribute equally to the frequency. In the Ashkenazi Jewish population, two genes drive the high frequency of pathogenic variants, with their variants achieving the same frequency – KCNQ1 (variant 11-2593251-G-A) and TTN (variant 2-179454576-G-A). The frequencies of these variants are stated in Table 4.4.

The LDLR (low-density lipoprotein receptor) gene family encodes cell surface proteins involved in receptor-mediated endocytosis of specific ligands. Mutations in this gene lead to familial hypercholesterolemia, an autosomal dominant disorder [40]. KCNQ1 (potassium voltage-gated channel subfamily Q member 1) encodes a voltage-gated potassium channel required for the repolarization phase of the cardiac action potential. Mutations in this gene are linked to hereditary long QT syndrome 1 (Romano-Ward syndrome) [41]. TTN (titin) encodes a large abundant protein found in striated muscle. Mutations in this gene are associated with familial hypertrophic cardiomyopathy [42].

Table 4.2: This table presents genes associated with cardiovascular phenotypes, describing their modes of inheritance and frequencies of pathogenic/likely pathogenic variants by each population. The populations are namely: Latino/Admixed American (amr), Ashkenazi Jews (asj), African/African American (afr), East Asian (eas), Finnish (fin), North-Western European (nwe), Southern European (seu), Estonian (est) and Other Non-Finnish European (onf). In the mode of inheritance column, *AD* refers to autosomal dominant, *AR* to autosomal recessive, and *SD* to semidominant. The mode of inheritance was incorporated from Miller et al. [27].

Disease Group and Gene	Mode of inheritance	Frequency of pathogenic variants								
		amr	asj	afr	eas	fin	nwe	seu	est	onf
Cardiovascular phenotypes										
ACTA2	AD	0.0	0.0	0.0	0.0	0.000288	0.0	0.0	0.0	0.0
ACTC1	AD	-	-	-	-	-	-	-	-	-
APOB	AD	0.0	0.0	0.000115	0.0	0.0	0.000931	0.0	0.000872	0.001403
BAG3	AD	-	-	-	-	-	-	-	-	-
CALM1	AD	-	-	-	-	-	-	-	-	-
CALM2	AD	-	-	-	-	-	-	-	-	-
CALM3	AD	-	-	-	-	-	-	-	-	-
CASQ2	AR	0.0	0.0	0.000115	0.0	0.0	0.000465	0.0	0.0	0.000468
COL3A1	AD	0.0	0.0	0.000115	0.0	0.0	0.0	0.0	0.0	0.0
DES	AD	0.0	0.0	0.0	0.0	0.0	0.000233	0.0	0.0	0.000468
DSC2	AD	0.0	0.0	0.000115	0.0	0.0	0.0	0.0	0.0	0.0
DSG2	AD	0.0	0.0	0.000231	0.0	0.0	0.0	0.0	0.0	0.0
DSP	AD	0.0	0.0	0.000345	0.0	0.0	0.0	0.0	0.0	0.0
FBN1	AD	0.0	0.0	0.000115	0.000641	0.000288	0.000233	0.0	0.000218	0.0
FLNC	AD	0.0	0.0	0.0	0.0	0.0	0.000233	0.0	0.0	0.0
KCNH2	AD	0.0	0.0	0.000115	0.0	0.000576	0.0	0.0	0.000218	0.0
KCNQ1	AD	0.0	0.003448	0.00023	0.0	0.000576	0.000233	0.0	0.000218	0.000936
LDLR	SD	0.002358	0.0	0.001263	0.002567	0.000576	0.001861	0.0	0.000218	0.004677
LMNA	AD	0.0	0.0	0.0	0.0	0.0	0.000116	0.0	0.0	0.000468
MYBPC3	AD	0.0	0.0	0.000919	0.001282	0.000576	0.001047	0.0	0.0	0.0
MYH7	AD	0.001179	0.0	0.000115	0.0	0.0	0.000814	0.0	0.000436	0.001871
MYH11	AD	-	-	-	-	-	-	-	-	-
MYL2	AD	-	-	-	-	-	-	-	-	-
MYL3	AD	-	-	-	-	-	-	-	-	-
PCSK9	AD	-	-	-	-	-	-	-	-	-
PKP2	AD	0.0	0.0	0.000574	0.000642	0.0	0.000582	0.0	0.0	0.001403
PRKAG2	AD	-	-	-	-	-	-	-	-	-
RBM20	AD	-	-	-	-	-	-	-	-	-
RYR2	AD	-	-	-	-	-	-	-	-	-
SCN5A	AD	0.0	0.0	0.000459	0.0	0.0	0.000466	0.0	0.0	0.0
SMAD3	AD	-	-	-	-	-	-	-	-	-

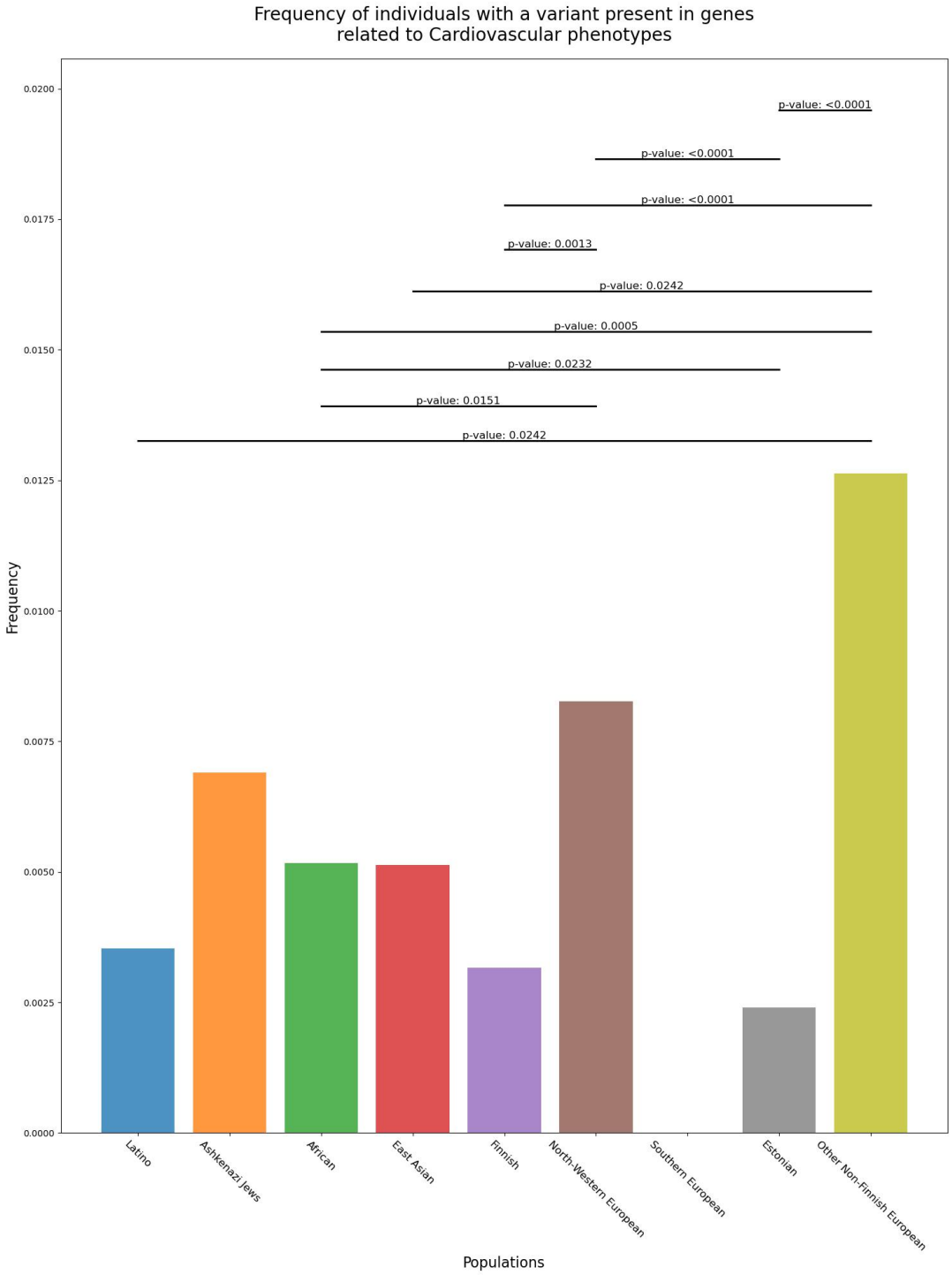

Figure 4.3: Bar plot showing the frequency of individuals carrying pathogenic/likely pathogenic variants in genes associated with cardiovascular phenotypes across different populations. These results are derived from the filtering step that excluded any variants classified as benign, likely benign, or of uncertain significance and variants not reported in the ClinVar database. The plot includes p-values that indicate the statistical significance of the differences in frequencies, derived from Fisher's Exact Test.

4.3 Inborn errors of metabolism phenotypes

This section presents the results of the analysis of variants from genes associated with inborn errors of metabolism phenotypes. Table 4.3 provides additional information regarding these genes. Due to the pathogenic mechanisms of diseases of inborn errors of metabolism phenotypes, we included only the variants reported in ClinVar as pathogenic/likely pathogenic. Genes associated with inborn errors of metabolism are typically autosomal recessive or X-linked. The individuals with these mutations and being reported in gnomAD are mostly carriers within the population [43].

Figure 4.4 shows the frequencies of individuals carrying pathogenic/likely pathogenic variants in genes associated with inborn errors of metabolism phenotypes across various populations. The Southern European population exhibits a high frequency of these variants, however, this should not be considered statistically significant. The second highest frequency is observed in the African/African American population, at just above 0.5%. This frequency diverges significantly from the ones in the Finnish and Estonian populations. The lowest frequency of pathogenic/likely pathogenic variants is exhibited in the Estonian population, which significantly diverges from the frequencies in the Finnish, North-Western European, and Southern European populations.

For the three populations with the highest frequencies of pathogenic/likely pathogenic variants, we again identified the gene and variant driving the high frequency. In the African/African American population, the gene driving the high frequency is GAA, with its variant 17-78092070-C-T. In the Other Non-Finnish European population, the gene is BTD, with its variant 3-15686731-A-C. For the East Asian population, it is also the GAA gene, but another variant, 17-78086721-C-A. The frequencies of these variants are stated in Table 4.4.

GAA (alpha glucosidase) encodes lysosomal alpha-glucosidase, which is essential for the degradation of glycogen to glucose in lysosomes. Mutations in this gene lead to glycogen storage disease II, also known as Pompe's disease, an autosomal recessive disorder [44]. BTD (biotinidase) encodes a protein that recycles protein-bound biotin by cleaving biocytin (biotin-epsilon-lysine), a byproduct of carboxylase degradation, regenerating free biotin. Mutations in this gene are linked to biotinidase deficiency [45].

Table 4.3: This table presents genes associated with cardiovascular, inborn errors of metabolism and miscellaneous phenotypes, describing their modes of inheritance and frequencies of pathogenic/likely pathogenic variants by each population. The populations are namely: Latino/Admixed American (amr), Ashkenazi Jews (asj), African/African American (afr), East Asian (eas), Finnish (fin), North-Western European (nwe), Southern European (seu), Estonian (est) and Other Non-Finnish European (onf). In the mode of inheritance column, *AD* refers to autosomal dominant, *AR* to autosomal recessive, and *XL* to X-linked. The mode of inheritance was incorporated from Miller et al. [27].

Disease Group and Gene	Mode of inheritance	Frequency of pathogenic variants								
		amr	asj	afr	eas	fin	nwe	seu	est	onf
Cardiovascular phenotypes continued										
TGFBR1	AD	-	-	-	-	-	-	-	-	-
TGFBR2	AD	-	-	-	-	-	-	-	-	-
TMEM43	AD	-	-	-	-	-	-	-	-	-
TNNC1	AD	-	-	-	-	-	-	-	-	-
TNNI3	AD	0.0	0.0	0.0	0.0	0.0	0.000116	0.0	0.0	0.0
TNNT2	AD	0.0	0.0	0.000115	0.0	0.0	0.0	0.0	0.0	0.0
TPM1	AD	-	-	-	-	-	-	-	-	-
TRDN	AR	-	-	-	-	-	-	-	-	-
TTN	AD	0.0	0.003448	0.00023	0.0	0.000288	0.000931	0.0	0.000218	0.000936
Inborn errors of metabolism phenotypes										
BTD	AR	0.003538	0.0	0.001493	0.0	0.0	0.001978	0.009434	0.000218	0.003274
GAA	AR	0.001179	0.003448	0.004367	0.005128	0.0	0.003142	0.009434	0.002183	0.001403
GLA	XL	-	-	-	-	-	-	-	-	-
OTC	XL	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000682
Miscellaneous phenotypes										
ACVRL1	AD	0.0	0.0	0.00046	0.0	0.0	0.000233	0.0	0.0	0.000468
ATP7B	AR	0.008255	0.006897	0.002066	0.010256	0.001726	0.0057	0.0	0.009592	0.005145
CACNA1S	AD	0.0	0.0	0.0	0.0	0.0	0.000233	0.0	0.000219	0.0
ENG	AD	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000219	0.0
HFE	AR	0.0	0.0	0.0	0.001284	0.0	0.0	0.0	0.0	0.0
HNF1A	AD	0.0	0.0	0.000115	0.0	0.0	0.0	0.0	0.0	0.0
RPE65	AR	0.002358	0.0	0.002296	0.000641	0.0	0.000931	0.0	0.000218	0.000468
RYR1	AD	0.0	0.003448	0.001149	0.000642	0.000864	0.001978	0.009434	0.001966	0.001871
TTR	AD	0.001179	0.0	0.017218	0.000643	0.0	0.000116	0.0	0.0	0.0

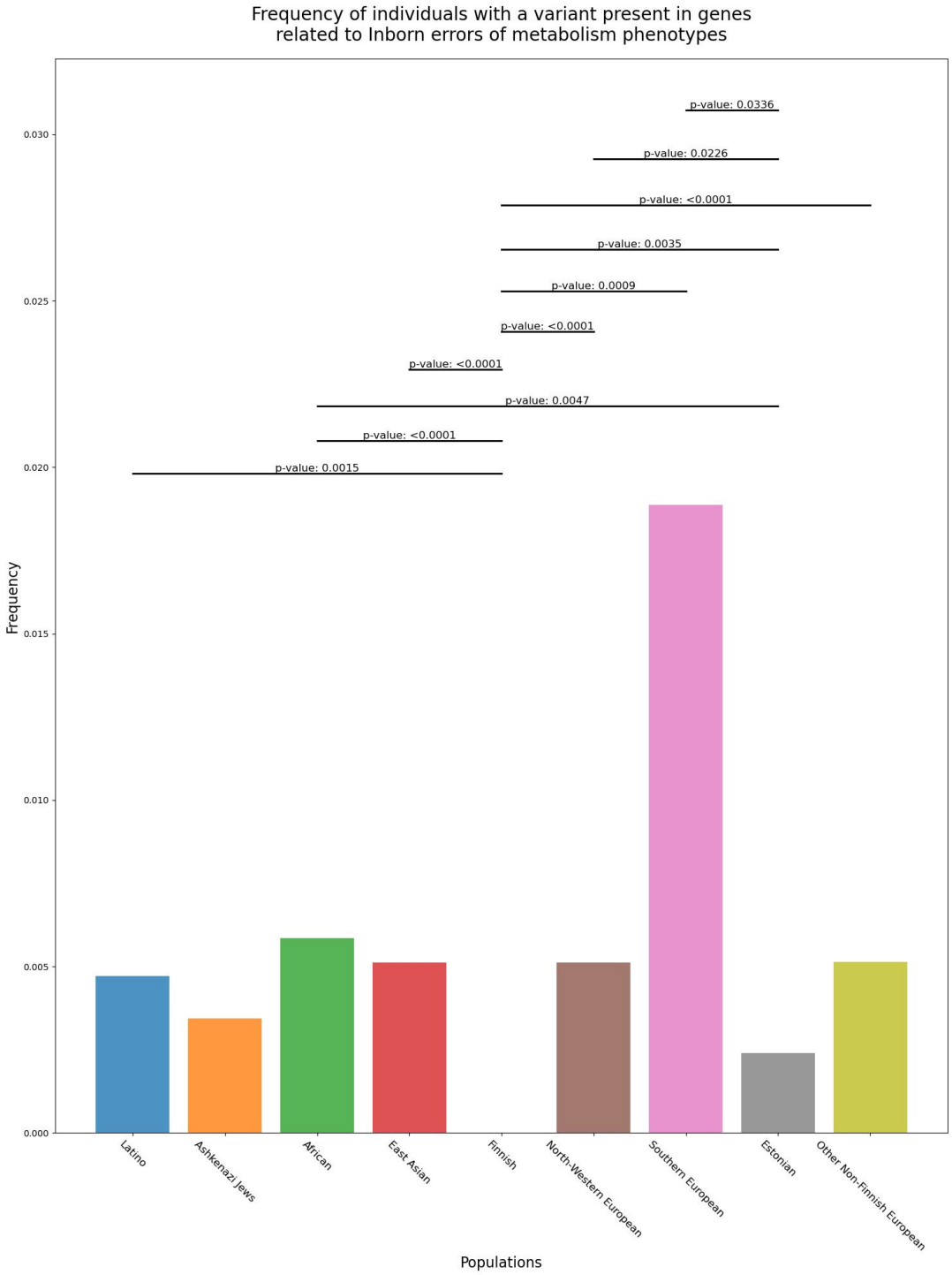

Figure 4.4: Bar plot showing the frequency of individuals carrying pathogenic/likely pathogenic variants in genes associated with inborn errors of metabolism phenotypes across different populations. These results are derived from the filtering step that excluded any variants classified as benign, likely benign, or of uncertain significance and variants not reported in the ClinVar database. The plot includes p-values that indicate the statistical significance of the differences in frequencies, derived from Fisher's Exact Test.

4.4 Miscellaneous phenotypes

This section presents the results of the analysis of variants from genes associated with miscellaneous phenotypes. Table 4.3 provides additional information regarding these genes. Due to the pathogenic mechanisms of diseases of the other phenotypes, we included only the variants reported in ClinVar as pathogenic/likely pathogenic.

Figure 4.5 illustrates the frequencies of individuals carrying pathogenic/likely pathogenic variants in genes associated with miscellaneous phenotypes across various populations. The population showing the highest frequency of these variants, with almost 2.5%, is the African/African American population. This frequency significantly diverges from the ones observed in the Latino/Admixed American, East Asian, Finnish, North-Western European, Estonian, and Other Non-Finnish European populations, with the highest level of significance (p-value below 0.0001) noted for Finnish, North-Western European, Estonian and Other Non-Finnish European populations. The lowest frequency is observed in the Finnish population.

For the three populations with the highest frequencies of pathogenic/likely pathogenic variants, we again identified the gene and variant driving the high frequency. In the African/African American population the gene driving high frequency is TTR, with its variant 18-29178618-G-A. In the East Asian population, the gene is ATP7B, with variant 13-52523908-G-C. In the Estonian population the gene is also ATP7B, but with another variant 13-52518281-G-T. The frequencies of these variants are stated in Table 4.4.

TTR (transthyretin) encodes a protein that transports thyroid hormones in the plasma and cerebrospinal fluid and is also involved in transporting retinol (vitamin A) in the plasma by binding with retinol-binding protein. Mutations in this gene are linked to amyloid deposition, primarily affecting peripheral nerves or the heart [46]. ATP7B (ATPase copper transporting beta) functions as a copper-transporting ATPase, exporting copper from cells. Mutations in this gene are associated with Wilson disease, characterized by copper accumulation [47].

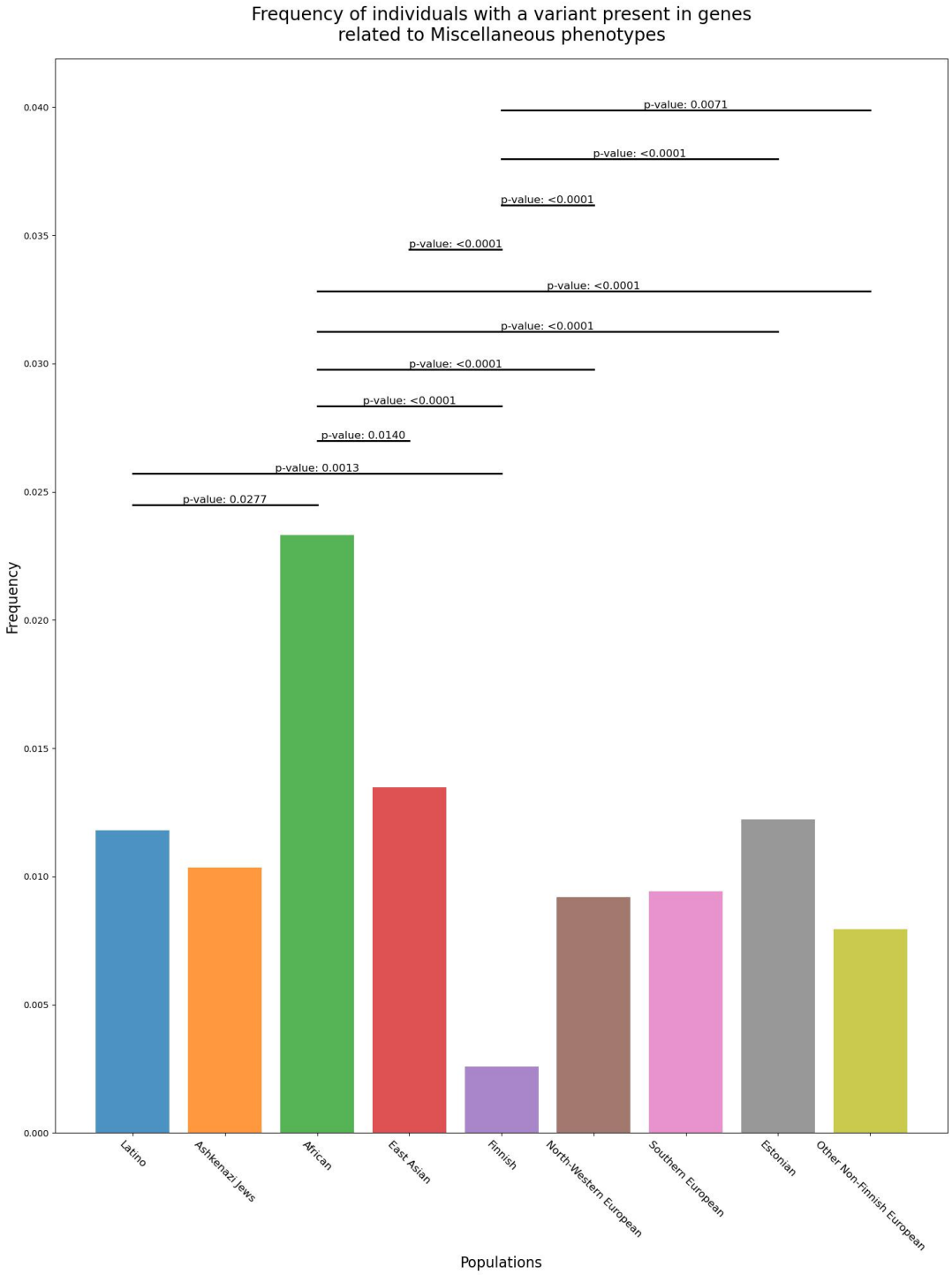

Figure 4.5: Bar plot showing the frequency of individuals carrying pathogenic/likely pathogenic variants in genes associated with miscellaneous phenotypes across different populations. These results are derived from the filtering step that excluded any variants classified as benign, likely benign, or of uncertain significance and variants not reported in the ClinVar database. The plot includes p-values that indicate the statistical significance of the differences in frequencies, derived from Fisher's Exact Test.

Table 4.4: This table lists the genes and variants driving the high frequency of pathogenic/likely pathogenic variants in the populations with the highest frequencies for each phenotype group. It includes the population name, gene name, position ID (chromosome-position-reference_allele-alternate_allele) of the variant within that gene, and frequency of that variant.

Population	Gene name	Position ID of the variant	Frequency
Cancer phenotypes (including high-impact variants not reported in ClinVar)			
Estonian	SDHD	11-111990013-A-T	0.00394
Ashkenazi Jews	SDHD	11-111990013-A-T	0.00885
Finnish	SDHD	11-111990013-A-T	0.010941
Cancer phenotypes (only variants reported in ClinVar)			
Ashkenazi Jews	BRCA2	13-32914437-GT-G	0.006897
Estonian	BRCA1	17-41243512-CT-C	0.001525
Other Non-Finnish European	PMS2	7-6022516-C-T	0.000478
Finnish	PALB2	16-23646274-CA-C	0.000864
Cardiovascular phenotypes			
Other Non-Finnish European	LDLR	19-11227604-G-A	0.001403
North-Western European	LDLR	19-11231156-G-A/19-11216133-G-A	0.000349/0.000349
Ashkenazi Jews	KCNQ1/TTN	11-2593251-G-A/2-179454576-G-A	0.003448/0.003448
Inborn errors of metabolism phenotypes			
African	GAA	17-78092070-C-T	0.001726
Other Non-Finnish European	BTD	3-15686731-A-C	0.000935
East Asian	GAA	17-78086721-C-A	0.003846
Miscellaneous phenotypes			
African	TTR	18-29178618-G-A	0.017218
East Asian	ATP7B	13-52523908-G-C	0.002564
Estonian	ATP7B	13-52518281-G-T	0.00873

Chapter 5

Discussion and conclusion

5.1 Discussion

In this research project, we compared the populational frequencies of pathogenic or likely pathogenic variants of ACMG-listed genes present in the gnomAD database. We attempted to identify differences in frequencies of these variants across different populations to find potential dependencies or correlations related to specific genes and also groups of genes associated with particular phenotypes – cancer, cardiovascular, inborn errors of metabolism, and miscellaneous phenotypes. Identification of the frequencies of pathogenic/likely pathogenic variants in different populations could allow us to understand population-specific genetic risks of various health conditions. This knowledge might be crucial to aid the development of targeted prevention and treatment strategies, and may also highlight the potential underrepresentation of certain populations in genetic studies.

The gnomAD database ensures a reliable source of genetic data because of quality control measures, including obtaining data primarily from case-control studies of common adult-onset diseases, removing samples with lower sequencing quality based on various metrics, and excluding samples from second-degree or closer related individuals across data types [17]. However, it lacks exhaustive clinical annotations or pathogenicity classifications. Hence, other databases such as ClinVar are necessary to cover this crucial information.

To analyze the gnomAD database, we first extracted the gene positions of the ACMG-listed genes from the full gnomAD database and manipulated the .vcf file format in order to reduce its size. Then, we annotated the database using VEP to add additional information to the variants that would aid the filtering process. Next, we filtered out the variants with frequency above 5%, which are considered benign according to ACMG criteria, since the population frequency is too high to cause a monogenic form of the disease [4]. Subsequently, we included only variants reported in ClinVar as pathogenic/likely pathogenic or variants with a 'HIGH'

impact according to VEP. A variant classified as having a 'HIGH' impact is presumed to disrupt significantly the function of a protein, likely causing protein truncation, loss of function, or triggering nonsense-mediated decay [48]. For data reproducibility, it is crucial to precisely track transcript isoforms and transcript versions. However, in some cases even including the version may not always prevent potential misinterpretations. Differences between a genome and a transcript set can lead to confusion and errors when reporting variants at the cDNA and genomic levels [23]. In the ClinVar database, variants are classified based on evidence, which is generally reliable, particularly, when the data is submitted by several submitters with no conflicts of interpretations [4].

Using the total (AN) and alternate (AC) numbers of alleles in samples, we calculated the frequencies of the given variants within each population. To calculate the frequencies of pathogenic/likely pathogenic variants per gene, we used the medians of AN values divided by sums of AC values per gene. Using the median of AN values ensures that extreme values are not affecting the calculation disproportionately and ensures that the calculations are representative of the typical population. Summing the alternate allele counts gives the total number of pathogenic variants observed across all samples, which provides insight into the overall burden of pathogenic variants in a given gene.

We also calculated the frequencies for four different phenotypes associated with the ACMG-listed genes, using medians of AN values and sums of AC values for all variants associated with one phenotype group. Finally, we performed Fisher's exact test to find significant differences in frequencies between populations for both individual genes and phenotype groups per population.

The results from Venner et al. [16] for the gnomAD database (compared with the All of Us cohort) are similar to our findings. However, they used a different, newer assembly of the human genome – GRCh38 – while in our research we used GRCh37. Hence, the data may differ slightly. For instance, in our study, the frequency of pathogenic/likely pathogenic variants in the BRCA2 gene in the African population is somewhat higher compared to results from Venner et al.; on the other hand, for the LDLR gene in the Latino/Admixed American population, the frequency is at the same level.

We based our methods on the approaches of Jensson et al. [2] and Venner et al. [16]. However, while Jensson et al. focused on one specific population, we analyzed data across different populations. Moreover, Venner et al. while using the gnomAD database focused on variants grouped by genes. In contrast, we added grouping by phenotypes to emphasize the diseases caused by genes in different populations and to highlight the importance of screenings and counseling for genes associated with specific diseases.

5.1.1 Cancer phenotypes

For genes associated with cancer phenotypes, we calculated the frequencies twice. In one experiment (CV_P+HIGH_VEP), we used variants reported in ClinVar as pathogenic/likely pathogenic as well as variants of high impact that were not necessarily reported in ClinVar. While some of these variants might have been reviewed with conflicting interpretations (with some classifiers potentially marking the variant as benign) or uncertain significance, therefore in the other experiment (CV_P), we included only variants that were reported in ClinVar as pathogenic/likely pathogenic. By including variants not reported in ClinVar in the CV_P+HIGH_VEP experiment, we incorporated a broader range of potentially significant variants, which could bring attention to variants that might be clinically relevant but not yet thoroughly reviewed. However, the CV_P experiment ensures that the analysis is based on variants with established clinical significance, which provides a more reliable assessment of the genetic risk associated with cancer phenotypes.

In cancer phenotypes, most genes follow a loss-of-function (LoF) mechanism, where the inactivation of tumor suppressor genes leads to cancer development [33, 34]. High-impact variants, even if not yet classified in the ClinVar database, can significantly disrupt gene function and potentially support disease progression [48]. As the mechanisms of cancer development are well-documented, analyzing both high-impact variants not included in ClinVar and only variants reported in ClinVar could provide a comprehensive view, and ensure a thorough and reliable analysis.

We observed both differences and similarities between the two experiments. When comparing the CV_P+HIGH_VEP test to the CV_P test, we reported that the frequencies of pathogenic/likely pathogenic variants dropped for all populations in the CV_P experiment. The Ashkenazi Jewish population retained a relatively high frequency, compared to other populations, in both experiments. Most other populations saw a drop of around 0.5%. However, the Estonian population experienced a significant decrease in the frequency of pathogenic/likely pathogenic variants in the CV_P experiment.

The differences observed in cancer phenotypes between the two experiments might have several causes. The first analysis included both pathogenic/likely pathogenic variants reported in ClinVar database, as well as high-impact variants not present in ClinVar. This broader set of potentially significant variants provides a larger variety of data, however, it might be not as reliable, while the variants lack the substantial evidence from research studies that is necessary for ClinVar submissions. As the less-documented variants in some populations were not included in the calculations, the frequencies of pathogenic/likely pathogenic variants were lower in the second analysis. Additionally, certain populations might be underrepresented in the genetic research and routine, which results in fewer submissions to ClinVar. Newly identified high-impact variants might not yet be recognized as clinically significant, also leading to

fewer submissions. These differences between plots emphasize the need for more personalized genetic research, that could increment the number of submissions to various databases and enhance the comprehensiveness and reliability of genetic variant data interpretation.

For the CV_P+HIGH_VEP experiment, the highest frequency of pathogenic/likely patho-genic variants was exhibited by the Estonian population. This frequency was significantly different from frequencies observed in the African/African American, East Asian, Finnish, North-Western European, and Other Non-Finnish European populations.

In the CV_P experiment, the highest frequency of pathogenic or likely pathogenic variants was exhibited by the Ashkenazi Jewish population. This frequency was significantly different from frequencies observed in the African/African American, East Asian, Finnish, North-Western European and Other Non-Finnish European populations.

The results from the CV_P experiment are supported by already existing research. The Ashkenazi Jewish population is known to have a high risk associated with BRCA genes, which are linked to cancer [13]. The Estonian population also has known pathogenicity risk of cancer phenotypes, particularly with the BRCA1 gene [49]. For the Finnish population, there is evidence of association with diseases caused by mutations in the PALB2 gene, however, BRCA genes are reported to be more frequent in this population according to the literature [14]. This differs from our results, as we did not observe any variants related to BRCA1 or BRCA2 genes in the Finnish population. Still, these findings are generally consistent with the obtained results. However, for the CV_P+HIGH_VEP experiment, there is no clear evidence of the SDHD gene being particularly frequent in any of the populations, which indicates the need for more thorough research.

5.1.2 Cardiovascular phenotypes

For genes associated with cardiovascular phenotypes, we calculated the frequencies of pathogenic or likely pathogenic variants only reported in the ClinVar database as pathogenic/likely pathogenic. This approach ensured that the analysis was based on variants with established clinical significance, providing a more reliable assessment of the genetic risk of these diseases. We included only these variants because of the pathogenic mechanisms of diseases of cardiovascular phenotypes. Genetic variants associated with cardiovascular phenotypes can exhibit different levels of effect and can be differently expressed in different individuals [50]. Due to that variability, it is more reliable to depend on reviewed variants that are confirmed as pathogenic/likely pathogenic.

The population with the highest frequency of pathogenic/likely pathogenic variants in genes associated with cardiovascular phenotypes is the Other Non-Finnish European population. This

frequency is significantly different from frequencies observed in the Latino/Admixed American, African/African American, East Asian, Finnish, and Estonian populations. According to some literature, Eastern European populations have a high risk of cardiovascular diseases, and other European populations are also at greater risk, especially given the populations are getting older [51].

5.1.3 Inborn errors of metabolism phenotypes

For genes associated with inborn errors of metabolism phenotypes, we calculated the frequencies of pathogenic/likely pathogenic variants only reported in the ClinVar database as pathogenic or likely pathogenic. This approach ensured that the analysis was based on variants with established clinical significance, providing a more reliable assessment of the genetic risk of these diseases. We focused on these specific variants because of the pathogenic mechanisms of diseases related to inborn errors of metabolism phenotypes. Most diseases caused by inborn errors of metabolism are inherited in an autosomal recessive manner, meaning that individuals must inherit two copies of the pathogenic variant to manifest the disease. These conditions often exhibit symptoms in the early stages of life [43]. Since the gnomAD database consists of data from individuals presumed to be healthy with potential predispositions to various diseases [52], we focused only on clinically validated pathogenic/likely pathogenic variants associated with genes related to inborn errors of metabolism. This approach allowed us to indicate these individuals as carriers of the potential disease and ensured a more reliable assessment of genetic risk for carriership. These results might emphasize the importance of offering future parents screening for carriership of pathogenic variants associated with this phenotype.

The population with the highest frequency of pathogenic/likely pathogenic variants in genes associated with inborn errors of metabolism phenotypes is the African/African American population. This frequency is significantly different from frequencies observed in the Finnish and Estonian populations. Literature suggests that high rates of inborn errors of metabolism diseases are exhibited also by the Eastern Mediterranean population [53].

5.1.4 Miscellaneous phenotypes

For genes associated with other phenotypes, we calculated the frequencies of pathogenic/likely pathogenic variants only from those reported in the ClinVar database as pathogenic/likely pathogenic. This approach ensured that the analysis was based on variants with established clinical significance to provide more reliable assessment of the genetic risk of these diseases. We included only ClinVar pathogenic/likely pathogenic variants because miscellaneous phenotypes can involve a variety of genetic mechanisms and using only the ClinVar-reported variants ensures consistency and reliability.

The population with the highest frequency of pathogenic/likely pathogenic variants in genes associated with miscellaneous phenotypes is the African/African American population. This frequency is significantly different from frequencies observed in the Latino/Admixed American, East Asian, Finnish, North-Western European, Estonian and Other Non-Finnish European populations. According to the literature, variants in the TTR gene are found among individuals of African/African American descent [54]. The high genetic diversity of individuals within the African population can lead to increased detection of genetic variants. Limited access to genetic screening in some African regions can contribute to misdiagnosis of various conditions, resulting in more diseases being classified with less-defined phenotypes [55].

In this project, we observed differences in the frequencies of pathogenic/likely pathogenic variants across different populations which highlights the genetic diversity among individuals and populations. This research could help identify population-specific genetic risks to perform personalized genetic screenings and inform clinical recommendations and counseling towards specific populations or individuals. Detailed information on population-specific variant frequencies can provide valuable insights for genetic counseling and personalization of genetic research. Moreover, this study highlights the underrepresentation of certain populations. For instance, European populations are divided into several categories, while the African/African American or Latino/Admixed American populations are grouped together as a single category.

5.2 Limitations and future work

The gnomAD genomic database contains extensive information on human genetic variants, which allowed for a comprehensive analysis of these variants across different populations. Using the gnomAD database ensured a variety of data from a wide range of populations, reducing bias towards particular populations. However, some populations might still be underrepresented, therefore other datasets could be incorporated, especially the ones focused on specific populations or genes. This could address the potential gaps and enhance the specificity of the research, particularly for the populations with the highest frequencies of pathogenic/likely pathogenic variants, to validate the results.

Further enhancement of this study could include incorporating the MANE (Matched Annotation from NCBI and EMBL-EBI) transcript in the VEP annotation to improve the classification and selection of the variants. The MANE transcript is designed to be a universally adopted standard reference for variant reporting, providing consistency across browser displays, resources, and tools. It ensures a high-value set of transcripts and corresponding proteins, essential for understanding the impact of clinically relevant variants. Moreover, the MANE Select set identifies a representative transcript for each human protein-coding gene, which includes transcripts for all ACMG-reported genes. Extensive adoption of these transcript sets can enhance consis-

tency in reporting, facilitate the data exchange regardless of annotation source, and streamline clinical interpretation [56]. Including the MANE transcript as a flag during the VEP annotation would establish that variants from different transcripts are distinctively marked, one being MANE, improving the overall utility of the data.

Additional data could be incorporated into the research by considering more recent gnomAD versions – i.e. the ones that are aligned to the GRCh38, more recent genome assembly, compared to GRCh37. The updates in GRCh38 make it a more robust foundation for comprehensive analyses [57]. Due to the time constraints of this project, we decided to focus this study only on the GRCh37 assembly, because the data regarding the GRCh38 assembly is even broader. The annotation and analysis processes were already time-consuming with GRCh37, so using the more extensive GRCh38 dataset would have taken even more time.

Another addition could involve integrating the Alpha Missense score. This score can be added as a plugin in VEP to then annotate missense variants with the pre-computed AlphaMissense pathogenicity scores [58]. AlphaMissense is a deep learning model developed by Google DeepMind that predicts the pathogenicity of single nucleotide missense variants. While only a small fraction of variants have been experimentally studied, there are extensive amounts of biological sequence data suitable to be the training data for machine learning models [59]. This integration would aid the classification of modifier variants with moderate impact.

5.3 Conclusion

In this research project, we identified differences in frequencies of pathogenic/likely pathogenic variants among different populations which highlights the genetic diversity between individuals and populations. As the pathogenic/likely pathogenic variants are reported to reduce lifespan [2], it is crucial to emphasize the need for personalized and population-based screening. We found that for cancer phenotypes the highest frequencies of pathogenic/likely pathogenic variants were observed in the Estonian and Ashkenazi Jewish populations. For cardiovascular phenotypes, the highest frequencies were in the Other Non-Finnish European and North-Western European populations. For inborn errors of metabolism phenotypes the African/African American and Other Non-Finnish European had the highest frequencies. For miscellaneous phenotypes the highest frequencies were found in the African/African American and Estonian populations.

If certain pathogenic variants are more common in specific populations, personalized medicine approaches can be developed to target these variants, leading to more effective treatments. Some variants may influence the drug metabolism of individuals, thus understanding the frequencies of such variants can help predict drug response and avoid negative reactions in different populations. Personalized medicine approaches should be prioritized more often, allowing

researchers and clinicians to work towards increasing the lifespan of individuals suffering or yet to be suffering from various diseases.

Bibliography

[1] Robert C Green, Jonathan S Berg, Wayne W Grody, Sarah S Kalia, Bruce R Korf, Christa L Martin, Amy L McGuire, Robert L Nussbaum, Julianne M O'Daniel, Kelly E Ormond, et al. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genetics in Medicine*, 15(7):565–574, 2013.

[2] Brynjar O Jensson, Gudny A Arnadottir, Hildigunnur Katrinardottir, Run Fridriksdottir, Hannes Helgason, Asmundur Oddsson, Gardar Sveinbjornsson, Hannes P Eggertsson, Gisli H Halldorsson, Bjarni A Atlason, et al. Actionable genotypes and their association with life span in Iceland. *New England Journal of Medicine*, 389(19):1741–1752, 2023.

[3] National Human Genome Research Institute. Iceland study provides insights into disease, paves way for large-scale genomic studies. https://www.genome.gov/27561444/iceland-study-provides-insights-into-disease-paves-way-for-largescale-genomic-studies.

[4] Sue Richards, Nazneen Aziz, Sherri Bale, David Bick, Soma Das, Julie Gastier-Foster, Wayne W Grody, Madhuri Hegde, Elaine Lyon, Elaine Spector, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in medicine*, 17(5):405–423, 2015.

[5] Yan Guo, Yulin Dai, Hui Yu, Shilin Zhao, David C Samuels, and Yu Shyr. Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis. *Genomics*, 109(2):83–90, 2017.

[6] EMBL-EBI. Ensembl – Genome assemblies. https://www.ensembl.org/info/genome/genebuild/assembly.html.

[7] AE Rougvie. Heterochronic Mutation. In *Brenner's Encyclopedia of Genetics: Second Edition*, pages 442–445. Elsevier Inc., 2013.

[8] Melissa J Landrum, Shanmuga Chitipiralla, Garth R Brown, Chao Chen, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Wonhee Jang, Kuljeet Kaur, Chunlei Liu, et al. ClinVar: improvements to accessing data. *Nucleic acids research*, 48(D1):D835–D844, 2020.

[9] Robert C Green, Jonathan S Berg, Wayne W Grody, Sarah S Kalia, Bruce R Korf, Christa L Martin, Amy L McGuire, Robert L Nussbaum, Julianne M O'Daniel, Kelly E Ormond, et al. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genetics in medicine*, 15(7):565–574, 2013.

[10] Lynn B Jorde. Genetic variation and human evolution. *American Society of Human Genetics*, 7(2019):28–33, 2003.

[11] Ken Batai, Stanley Hooker, and Rick A Kittles. Leveraging genetic ancestry to study health disparities. *American Journal of Physical Anthropology*, 175(2):363–375, 2021.

[12] Deepti Gurdasani, Inês Barroso, Eleftheria Zeggini, and Manjinder S Sandhu. Genomics of disease risk in globally diverse populations. *Nature Reviews Genetics*, 20(9):520–535, 2019.

[13] Ellen Warner, William Foulkes, Pamela Goodwin, Wendy Meschino, John Blondal, Colleen Paterson, Hilmi Ozcelik, Paul Goss, Diane Allingham-Hawkins, Nancy Hamel, et al. Prevalence and penetrance of BRCA1 and BRCA2 gene mutations in unselected Ashkenazi Jewish women with breast cancer. *Journal of the National Cancer Institute*, 91(14):1241–1247, 1999.

[14] Anna K Nurmi, Maija Suvanto, Joe Dennis, Kristiina Aittomäki, Carl Blomqvist, and Heli Nevanlinna. Pathogenic Variant spectrum in breast cancer risk genes in Finnish patients. *Cancers*, 14(24):6158, 2022.

[15] Laura H Goetz and Nicholas J Schork. Personalized medicine: motivation, challenges, and progress. *Fertility and sterility*, 109(6):952–963, 2018.

[16] Eric Venner, Karynne Patterson, Divya Kalra, Marsha M Wheeler, Yi-Ju Chen, Sara E Kalla, Bo Yuan, Jason H Karnes, Kimberly Walker, Joshua D Smith, et al. The frequency of pathogenic variation in the All of Us cohort reveals ancestry-driven disparities. *Communications biology*, 7(1):174, 2024.

[17] Konrad J Karczewski, Laurent C Francioli, Grace Tiao, Beryl B Cummings, Jessica Alföldi, Qingbo Wang, Ryan L Collins, Kristen M Laricchia, Andrea Ganna, Daniel P Birnbaum, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581(7809):434–443, 2020.

[18] gnomAD Production Team. gnomAD – v2 Downloads. https://gnomad.broadinstitute.org/downloads#v2.

[19] Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A Albers, Eric Banks, Mark A DePristo, Robert E Handsaker, Gerton Lunter, Gabor T Marth, Stephen T Sherry, et al. The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158, 2011.

[20] Heng Li. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21):2987–2993, 2011.

[21] Sanger Institute. bcftools(1) Manual Page. https://samtools.github.io/bcftools/bcftools.html.

[22] Petr Danecek, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, Thomas Keane, Shane A McCarthy, Robert M Davies, et al. Twelve years of SAMtools and BCFtools. *Gigascience*, 10(2):giab008, 2021.

[23] William McLaren, Laurent Gil, Sarah E Hunt, Harpreet Singh Riat, Graham RS Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. The ensembl variant effect predictor. *Genome biology*, 17:1–14, 2016.

[24] Melissa J Landrum, Jennifer M Lee, George R Riley, Wonhee Jang, Wendy S Rubinstein, Deanna M Church, and Donna R Maglott. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic acids research*, 42(D1):D980–D985, 2014.

[25] NCBI. ClinVar FTP site (/pub/clinvar). https://ftp.ncbi.nlm.nih.gov/pub/clinvar/.

[26] NCBI. ClinVar FTP site – README file for ftp://ftp.ncbi.nih.gov/pub/clinvar. https://ftp.ncbi.nlm.nih.gov/pub/clinvar/tab_delimited/README.

[27] David T Miller, Kristy Lee, Noura S Abul-Husn, Laura M Amendola, Kyle Brothers, Wendy K Chung, Michael H Gollob, Adam S Gordon, Steven M Harrison, Ray E Hershberger, et al. ACMG SF v3. 2 list for reporting of secondary findings in clinical exome and genome sequencing: A policy statement of the American College of Medical Genetics and Genomics (ACMG), 2023.

[28] NCBI. ACMG Recommendations for Reporting of Secondary Findings in Clinical Exome and Genome Sequencing. https://www.ncbi.nlm.nih.gov/clinvar/docs/acmg/.

[29] Free Software Foundation. Bash Reference Manual. https://www.gnu.org/software/bash/manual/bash.html.

[30] UCSC Genome Browser. Frequently Asked Questions: Data File Formats – BED format. https://genome.cse.ucsc.edu/FAQ/FAQformat.html#format1.

[31] Tatsuaki Kurosaki and Lynne E Maquat. Nonsense-mediated mRNA decay in humans at a glance. *Journal of cell science*, 129(3):461–467, 2016.

[32] Fred R.T. Nelson and Carolyn Taliaferro Blauvelt. Chapter 13 - Introduction: The Research Enterprise. In Fred R.T. Nelson and Carolyn Taliaferro Blauvelt, editors, *A Manual of Orthopaedic Terminology (Seventh Edition)*, pages 353–369. Mosby, Philadelphia, seventh edition edition, 2007.

[33] Margaret P Adam, Jerry Feldman, Ghayda M Mirzaa, Roberta A Pagon, Stephanie E Wallace, Lora JH Bean, Karen W Gripp, and Anne Amemiya. GeneReviews®[Internet]. 1993.

[34] Johns Hopkins University. OMIM – An Online Catalog of Human Genes and Genetic Disorders. https://www.omim.org/.

[35] Madeline Foley, Anu Sharma, Kinley Garfield, Luke Maese, Luke Buchmann, Julie Boyle, Wendy Kohlmann, Joanne Jeter, and Samantha Greenberg. A need to tailor surveillance based on family history: describing a highly penetrant familial paraganglioma kindred with an SDHD pathogenic variant. *Familial Cancer*, 22(2):217–224, 2023.

[36] NCBI Gene Database. BRCA1 BRCA1 DNA repair associated [Homo sapiens (human)]. https://www.ncbi.nlm.nih.gov/gene/672.

[37] NCBI Gene Database. BRCA2 BRCA2 DNA repair associated [Homo sapiens (human)]. https://www.ncbi.nlm.nih.gov/gene/675.

[38] NCBI Gene Database. PMS2 PMS1 homolog 2, mismatch repair system component [Homo sapiens (human)]. https://www.ncbi.nlm.nih.gov/gene/5395.

[39] NCBI Gene Database. PALB2 partner and localizer of BRCA2 [Homo sapiens (human)]. https://www.ncbi.nlm.nih.gov/gene/79728.

[40] NCBI Gene Database. LDLR low density lipoprotein receptor [Homo sapiens (human)]. https://www.ncbi.nlm.nih.gov/gene/3949.

[41] NCBI Gene Database. KCNQ1 potassium voltage-gated channel subfamily Q member 1 [Homo sapiens (human)]. https://www.ncbi.nlm.nih.gov/gene/3784.

[42] NCBI Gene Database. TTN titin [Homo sapiens (human)]. https://www.ncbi.nlm.nih.gov/gene/7273.

[43] Ana Maria Martins. Inborn errors of metabolism: a clinical overview. *Sao Paulo Medical Journal*, 117:251–265, 1999.

[44] NCBI Gene Database. GAA alpha glucosidase [Homo sapiens (human)]. https://www.ncbi.nlm.nih.gov/gene/2548.

[45] NCBI Gene Database. BTD biotinidase [Homo sapiens (human)]. https://www.ncbi.nlm.nih.gov/gene/686.

[46] NCBI Gene Database. TTR transthyretin [Homo sapiens (human)]. https://www.ncbi.nlm.nih.gov/gene/7276.

[47] NCBI Gene Database. ATP7B ATPase copper transporting beta [Homo sapiens (human)]. https://www.ncbi.nlm.nih.gov/gene/540.

[48] National Cancer Institute at the National Institutes of Health. In the Most Frequent Mutations table for the VEP impact score, which algorithm in the VEP is the GDC using to determine "H" or "M"? https://gdc.cancer.gov/content/most-frequent-mutations-table-vep-impact-score-which-algorithm-vep-gdc-using-determine-

[49] Kristiina Tamboom, Krista Kaasik, Jelena Aršavskaja, Mare Tekkel, Aili Lilleorg, Peeter Padrik, Andres Metspalu, and Toomas Veidebaum. BRCA1 mutations in women with familial or early-onset breast cancer and BRCA2 mutations in familial cancer in Estonia. *Hereditary cancer in clinical practice*, 8:1–7, 2010.

[50] David Seo, Geoffrey S Ginsburg, and Pascal J Goldschmidt-Clermont. Gene expression analysis of cardiovascular diseases: novel insights into biology and clinical applications. *Journal of the American College of Cardiology*, 48(2):227–235, 2006.

[51] Daan Kromhout. Epidemiology of cardiovascular diseases in Europe. *Public health nutrition*, 4(2b):441–457, 2001.

[52] Konrad J Karczewski, Ben Weisburd, Brett Thomas, Matthew Solomonson, Douglas M Ruderfer, David Kavanagh, Tymor Hamamsy, Monkol Lek, Kaitlin E Samocha, Beryl B Cummings, et al. The ExAC browser: displaying reference data information from over 60 000 exomes. *Nucleic acids research*, 45(D1):D840–D845, 2017.

[53] Donald Waters, Davies Adeloye, Daisy Woolham, Elizabeth Wastnedge, Smruti Patel, and Igor Rudan. Global birth prevalence and mortality from inborn errors of metabolism: a systematic analysis of the evidence. *Journal of global health*, 8(2), 2018.

[54] Pranav Chandrashekar, Laith Alhuneafat, Meghan Mannello, Lana Al-Rashdan, Morris M Kim, Jason Dungu, Kevin Alexander, and Ahmad Masri. Prevalence and outcomes of p. Val142Ile TTR amyloidosis cardiomyopathy: a systematic review. *Circulation: Genomic and Precision Medicine*, 14(5):e003356, 2021.

[55] Luisa Pereira, Leon Mutesa, Paulina Tindana, and Michèle Ramsay. African genetic diversity and adaptation inform a precision medicine agenda. *Nature Reviews Genetics*, 22(5):284–306, 2021.

[56] Joannella Morales, Shashikant Pujar, Jane E Loveland, Alex Astashyn, Ruth Bennett, Andrew Berry, Eric Cox, Claire Davidson, Olga Ermolaeva, Catherine M Farrell, et al. A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. *Nature*, 604(7905):310–315, 2022.

[57] Tina Graves-Lindsay, Derek Albracht, Robert S Fulton, Milinn Kremitzki, Vincent Magrini, Chris Markovic, Sean McGrath, Karyn Meltz Steinberg, Richard K Wilson, et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Research*, 2017.

[58] EMBL-EBI. Ensembl Variant Effect Predictor – Plugins. https://www.ensembl.org/info/docs/tools/vep/script/vep_plugins.html.

[59] Jun Cheng, Guido Novati, Joshua Pan, Clare Bycroft, Akvilė Žemgulytė, Taylor Applebaum, Alexander Pritzel, Lai Hong Wong, Michal Zielinski, Tobias Sargeant, et al. Accurate

proteome-wide missense variant effect prediction with AlphaMissense. *Science*, 381(6664):eadg7492, 2023.