DEPARTMENT OF ENGINEERING MATHEMATICS

Multivariate analysis of genome-wide data of systolic blood pressure, diastolic blood pressure and pulse rate to identify pleiotropic genes for cardiovascular disease using MetaCCA

Gurdeep Singh Bhambra

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of Master of Science in the Faculty of Engineering.

Monday 12th September, 2022

Supervisor: Dr Colin Campbell

Declaration

This dissertation is submitted to the University of Bristol in accordance with the requirements of the degree of MSc in the Faculty of Engineering. It has not been submitted for any other degree or diploma of any examining body. Except where specifically acknowledged, it is all the work of the Author.

Gurdeep Singh Bhambra, Monday 12th September, 2022

Contents

List of Figures

List of Tables

List of Algorithms

Ethics Statement

This project did not require ethical review, as determined by my supervisor, Dr Colin Campbell.

Abstract

Cardiovascular disease is a term used to denote other diseases including cardiac muscle diseases and vascular system diseases. Cardiovascular disease is a significant cause of deaths in UK[23][3]. These diseases often cause mutations in specific regions of our DNA known as genes. Finding mutations in genes caused by these diseases or vice versa can help understand the disease better. Genome-wide association studies (GWAS) data often help in finding linear associations in diseases. Using canonical correlation analysis, one can find such associations from diseases. The summary results of GWAS are often anonymous which gives enough flexibility to handle it but it also cannot be applied to conventional canonical correlation analysis methods. To overcome this, the metaCCA model is used.

The aim of this thesis is to find common genes among systolic blood pressure, diastolic blood pressure and pulse rate traits related to cardiovascular diseases using metaCCA. Blood pressure and pulse rate are phenotypes directly related to the performance of the heart. The genes obtained from metaCCA are further analysed by using gene-set analysis, pathway enrichment analysis and GO enrichment analysis.

The achievements of this thesis are as follows:

1. R, python and bash script were used together throughout this analysis with many other command line tools such as plink[37] and online tools such as atlas[49] and enrichr[51][20][5].

2. The gene-set analysis and metaCCA phenotype correlation matrix confirmed the relationship between systolic blood pressure and diastolic blood pressure.

3. The pathways with the most significant overlapped genes were related to cardiac-related functions and diseases.

4. GO terms confirm that most of the genes obtained from metaCCA are related to cardiac biological processes, molecular functions and cellular components.

GitHub Repository: https://github.com/GurdeepSinghBhambra/MSc-Thesis-metaCCA-for-CVD.git
Oral Presentation: https://youtu.be/8ykT_TCwE8o

Supporting Technologies

- I used pandas, seaborn, matplotlib and dask python libraries to manipulate, process and visualize large datasets.
- Tidyverse, Bioconductor and metaCCA[7] R packages were used to apply the metaCCA model.
- I used bash script with tools such as plink[37], awk, and bcftools to manipulate and process large datasets with limited resources.
- online tools such as atlas[49] and enrichr[51][20][5] was used to do major part of the analysis.
- I used jupiter and markdown notebooks for python and R respectively.
- LaTeX to format my thesis was used, via the online service *Overleaf*.

Notation and Acronyms

CCA	:	Canonical Correlation
CVD	:	Cardiovascular Disease
DNA	:	Deoxyribonucleic Acid
GT	:	Genotype Count
GO	:	Gene Ontologies[2][1]
GWAS	:	Genome-Wide Association Study
LD	:	Linkage Disequilibrium
MAF	:	Minor Allele Frequency
NHS	:	National Health Service
SNP	:	Single Nucleotide Polymorphism
	⋮	
r^2	:	Linkage disequilibrium coefficient
S_{XX}	:	genotype-genotype correlation matrix
S_{YY}	:	phenotype-phenotype correlation matrix
S_{YY}	:	genotype-phenotype data with phenotype coefficients

Acknowledgements

I would like to thank my thesis supervisor Dr Colin Campbell in the department of mathematics at the University of Bristol. He helped me navigate through the project and helped me with many concepts outside the realm of data science. I also want to thank him for providing me access to additional computational resources and helping with accessing the UK BioBank. He also made sure I got all types of other resources required for my thesis, for which I am grateful to him.

I would also like to thank Dr Amy Osborne in the department of mathematics at the University of Bristol. She took time off her busy schedule to help me understand the technical concepts for my thesis.

Finally, I want to thank my family and friends for helping me navigate through this thesis. Without their support, it would have been extremely difficult for me to complete my thesis.

Chapter 1

Introduction

1.1 Effects of Cardiovascular Disease

In 2002, 40% of all deaths in UK was caused by cardiovascular diseases (CVD)[23]. About half of the 40% deaths were due to coronary artery disease[23]. In 2004, CVD cost £29.1 billion to the UK economy[23]. Health care was the major cost component of CVD which was about 60% of the total cost[23]. To reduce deaths due to CVDs, in 1999, the UK government took initiatives to lower CVD cases by 2010[23].

A decade later situation improved but it still was a major contributor to deaths. CVD was still the most common cause of death in women in the UK[3]. It was the second most common cause of death in men[3]. In England, NHS spent around £6.8 billion treating CVD in 2012/2013[3]. CVD still is a major contributor to deaths worldwide[14].

1.2 Cardiovascular Disease, Blood Pressure and Pulse Rate

Cardiovascular diseases are diseases related to various critical functions or organs of the body including cardiac muscle diseases and diseases related to the vascular system supplying the heart, and other vital organs[14]. This is an umbrella term for many other diseases[14]. The most common diseases under CVD are ischemic heart disease, stroke and congestive heart failure[14].

Blood pressure is the pressure of blood circulating in the body against the walls of different components of the circulatory system. The heart pumps most of this blood. Blood pressure is often measured with 2 types of pressure, systolic blood pressure and diastolic blood pressure. The pressure created due to the pumping of the heart is known as systolic blood pressure. The resistance created by different blood vessels against the flow of blood is known as diastolic blood pressure. Pulse rate is the rate at which our heart pumps blood. It is often used to monitor our heart performance and can tell a lot about it as well. Blood pressure and pulse rate have known relationship with cardiovascular disease[13][14].

1.3 Association between Genes and Diseases

Most diseases have some kind of traits associated with them. These traits could be observable changes or something which is not visible. With advancements in technology and the rise of new ways to understand our genome data, we can analyse changes in our bodies at the DNA level. These traits can be analysed to find new associations or regions in our DNA for a disease. Genes, in our DNA, take part in biological functions or components which can make changes in our body, including phenotypes[21]. Variation or alteration in genes can lead to changes in phenotypes.

One way to find these variations is through genome-wide association study (GWAS). GWAS data can help record and analyse the changes at gene level. Analysing GWAS can help understand the association of genes with diseases[27]. GWAS summary statistics further summarise the data and is anonymous. This anonymity gives a lot of flexibility to access GWAS data. GWAS summary data can be posted online and in some cases can also be used without much ethical complications.

1.4 Analysis to Understand the Genes

To analyse the data and find associations, a known method is canonical correlation analysis (CCA)[15]. Canonical correlation with GWAS can be used for analysis. Given the anonymity of GWAS, many of its features are often absent which makes it difficult to use CCA. metaCCA solves this by using an estimation method that uses a shrinkage algorithm[7]. metaCCA would give sets of genes that might be statistically significant but will need to be analysed by understanding the context of the genes. To understand this, I used gene-set analysis, pathway enrichment analysis and GO enrichment analysis.

From these analyses, the aim of the thesis is to find genes that are associated with cardiovascular disease. These genes can further be analysed to understand how they affect different biological processes and how they are related to the disease.

Chapter 2

Background

Context

This thesis focuses on many bioinformatics and meta-analysis concepts. This chapter elaborates on these concepts and explores relevant literature revolving around the topic. This section also tries to outline the reasoning, structure and expectations from the analysis.

2.1 Genes and Phenotypes

DNA in our body, stores sets of instructions in our cells[21]. These instructions help us to reproduce, grow and survive. DNA has some regions that have instructions to perform a function in our body, these regions are called genes[21]. Genes carry information from parents to offspring made up of nucleotides[21]. These sets of regions that are composed of instructions and delivered from parents to offspring are called genotypes[21]. Protein-coding genes and non-coding are types of molecular genes[35]. The instructions of the genes are used to create functional gene products during a process known as gene expression[10]. These gene products at the very end of gene expression affect phenotypes as the gene products are used in further processes[10].

Phenotypes are physical or visible traits of the body[21]. These traits can often be measured in some way[21]. These traits can be the color of an eye, skin color, or like the topic of this thesis, measurable physical traits like blood pressure. Phenotypes are the result of interaction between genotypes (predetermined set of instructions) and environment[21]. In the context of disease-induced traits, sometimes phenotypes can be associated with some particular diseases[39]. Some examples of these traits due to diseases can be inflammation, irregularities in pulse, etc.

A physical trait or abnormality in a person can be predicted using a genotype and phenotype correlation, given a bunch of the same mutations[30]. Mutations (changes in organisms due to changes in the chemical structure of DNA[21]) can affect genes which in turn can change their sequence, resulting in slightly different forms of genes, these are called alleles[21]. These mutations together can further change the phenotypes too[21].

Now, let's relate our understanding to the focus of this thesis. In this thesis, I focus on finding pleiotropic genes using blood pressure and pulse rate phenotypes for cardiovascular diseases. Pleiotropic genes are genes that are associated with more than one trait[21]. Phenotypes or traits are: systolic blood pressure, diastolic blood pressure, and pulse rate. It is important to note that pulse rate is not exactly heart rate. Pulse is the sensation of a heartbeat.

2.2 metaCCA

To analyse the statistical summary data, I use a computational framework named metaCCA[7]. To understand how this model works, it's important to first understand the data as well.

2.2.1 Genome-Wide Association Studies (GWAS)

There are many types of studies focused on associating variation in our DNA with diseases. This variation of a single nucleotide in the genome for any two humans is termed single-nucleotide polymorphism

(SNP)[21]. The idea is to try to associate these SNP variations with disease traits. Even though SNPs might not have a major impact on biological functions, they do change amino acids, and mRNA leading to functional consequences[50].

The genomic data for a disease or a trait is collected through a type of observational study, called genome-wide association study (GWAS). In a GWAS, many variations of phenotypes are associated with hundreds to millions of individuals, which further associate these variations with diseases[26]. These individuals taking part in such a study can either be someone affected by the disease or not. GWAS helps in recording the human genome across all the individuals involved in the study and further, we can use this data to understand the variations for diseases.

Summary statistics of GWAS are often the aggregated results of the study[27][24]. These summary statistics are often helpful to anonymize data which avoids the data being used to track an individual and also helps with sharing the results from the study[27]. The aggregated results often try to find associations of SNPs and traits which might help to understand biological functions[27].

Some examples of impactful GWAS studies includes identifying SNPs associated with psychiatric traits[27], effects of kidney functions on blood pressure[53] and discovering new regions in chromosomes for type-2 diabetes[25].

2.2.2 metaCCA Framework

When it comes to free and unrestricted access to GWAS data, summary statistics are often freely available on the internet as they are anonymous and thus can be shared easily[7][27]. metaCCA framework performs analyses of SNPs with multiple traits using canonical correlation, given the GWAS summary statistics data[7]. The GWAS summary statistics need to be univariate GWAS summary statistics with linear regression as the association predictive model[7].

metaCCA combines two previously known types of multivariate analysis[7]. The types of analysis are:

1. Association analysis of multiple traits and a single SNP[7]: In this type of analysis, one genetic variant or SNP is tested against multiple phenotypes. This helps to analyse the association of a single genetic variant with these traits and to understand whether it's associated with one or more than one traits[45]. The best SNP is often selected based on how strong the statistical association is found to be within a region in the genome[52].

2. Association analysis of a single trait and multiple SNPs[7]: In this type of analysis, the cause or effect of a trait is analysed jointly and conditionally from multiple SNPs[52].

metaCCA further adapts the canonical correlation analysis of multiple SNPs against multiple phenotypes for GWAS summary statistics data[7].

Canonical correlation analysis (CCA), prior to metaCCA was a known technique to discover linear relationships between 2 variables with GWAS data[7]. For any two vectors of random variables, if there exists some kind of correlation between these variables, then, canonical correlation analysis tries to find a linear relationship with a combination of these two variables which maximizes the correlation between these variables[15]. metaCCA goes a bit further to adapt CCA for GWAS summary statistics data[7].

In univariate GWAS analysis, a linear model like linear regression can be used to find the regression coefficient of a genotype by predicting the phenotype[7]. This coefficient can help to understand how the phenotype is affected by the genotype[7]. The inputs for metaCCA use these coefficients to calculate the covariance matrix between genotypes and phenotypes[7]. There are essentially three required input components to the metaCCA framework[7]:

1. Covariance matrix of genotypes and phenotypes consisting of regression coefficients calculated from univariate GWAS (S_{XY})[7].

2. Genotype-Genotype correlation matrix consisting of correlation between each SNPs (S_{XX})[7].

3. Phenotype correlation matrix between all the traits in a study (S_{YY})[7].

We can summarise the inputs as follows:

If β_{gp} denotes the regression coefficient of a genotype for a phenotype, then S_{XY} is calculated as follows[7].

$$S_{XY} = \begin{bmatrix} \beta_{11} \dots & \beta_{1p} \\ \beta_{21} \dots & \beta_{2p} \\ \vdots & \\ \beta_{g1} \dots & \beta_{gp} \end{bmatrix} \tag{2.1}$$

While applying the GWAS data, each row in the S_{XY} matrix would be the overlapping SNP with its reference data used to calculate S_{XX} and each column would be a unique trait in a GWAS study.

As discussed previously under section 2.1, genotypes are instruction sets passed from parents to offspring. Factors like mutation and recombination (the process of reproduction where DNA of both parents are combined[21]), combined with demographic effects like population can give rise to genetic variation, altering their structures[7]. Hence, it's important to calculate the genotype-genotype correlation matrix based on reference data of a specific population[7].

For the common SNPs between the reference data and the GWAS data, a genotypic correlation matrix (S_{XX}) is calculated using a reference population such as 1000-genome project consortium data[7].

To calculate the correlation between each trait or phenotype (S_{YY}), the regression coefficients from the univariate GWAS are used[7]. Using Pearson correlation, the column vector of each phenotype in S_{XY} is used with other phenotypes in the same matrix to calculate the correlation between each pair of traits[7]. It is also important to note that, the error decreases if more genetic variables are used as suggested in the metaCCA paper[7]. Hence, S_{XY} used in this stage should have all the variables in the original GWAS summary dataset.

To calculate the canonical correlation coefficient we use the following formula[7][15]:

$$r = \frac{a^T S_{XY} b}{\sqrt{a^T S_{XX} a}\sqrt{b^T S_{YY} b}} \tag{2.2}$$

In this formula, r is the maximum canonical correlation between the X and Y variables. a and b are the vectors that maximise the canonical correlation, r.

From the three matrices (S_{XX}, S_{XY} and S_{YY}), the resultant covariance matrix becomes[7]:

$$\sum = \begin{pmatrix} S_{XX} & S_{XY} \\ S_{XY}^T & S_{YY} \end{pmatrix} \tag{2.3}$$

The covariance matrix resultant is not always a positive semidefinite (PSD), which cannot be directly used with CCA (2.2)[7]. To convert the resultant matrix ($\sum$) which will satisfy this property, an iterative shrinkage algorithm is used[7]. Using this shrinkage algorithm, a nearest valid covariance is found[7]. The algorithm tries to iteratively shrink the off-diagonal values to zero[7]. It continues to do so until $\sum$ becomes positive semidefinite[7]. The pseudo-code of the algorithm as per the metaCCA paper[7] is as follows:

while $\sum$ *not PSD* **do**

$$\sum = 0.999 \times \sum \ ;$$

$$diag(\sum) = 1 \ ;$$

end

Algorithm 2.1: Shrinkage algorithm used in metaCCA[7]

The result after shrinkage is now applied to the CCA equation(2.2). metaCCA model can use one or more GWAS summary statistics data for multivariate-analysis[7]. Out of the two variants of metaCCA, I used the one that tests a set of genetic variants (SNPs) with a set of phenotypes[7]. This variant of metaCCA would require the S_{XX} matrix[7].

Before we start with the next section, let's recall the key points discussed till now. As stated earlier, genes have encoded sets of instructions that perform certain functions. These genes can be altered at some positions (known as SNPs) and can give rise to different traits called phenotypes. In this thesis, I aim to find pleiotropic genes (genes that affect more than one trait) for cardiovascular diseases using blood pressure and pulse rate phenotypes. I find these pleiotropic genes using canonical correlation analysis with a suitable GWAS summary statistic dataset. Since GWAS summary statistics datasets are aggregated results of the study, I simply cannot use the data with CCA. To analyse the GWAS summary

data using CCA, I use a computational framework named metaCCA. I test the sets of genotypes reference data against sets of the 3 phenotypes using metaCCA.

2.3 Multivariate Analysis using metaCCA

There are numerous papers that have identified genetic variants that were associated with different diseases using metaCCA[6][22][17][18]. Some of the diseases that are often analysed and have relatively successful discoveries of pleiotropic genes are diabetes, psychiatric disorders, obesity, heart diseases and kidney diseases[6][22][17][18]. metaCCA gives statistically related genes with all the phenotypes. Even though there might be a correlation sometimes these genes have to be refined further to have a handful of them which makes it easier to analyse in-context of bio-informatics. Let's explore the actual structure of carrying out this analysis based on papers[6][22][17][18] that used metaCCA for multivariate analysis.

Firstly, the three input components of metaCCA S_{XX}, S_{XY}, and S_{YY} matrices are prepared as discussed in section 2.2.2. The GWAS summary statistics data should have mandatory columns for SNPs, the base-pair position specifying where exactly the exact SNP is located within a chromosome, the chromosome number, the regression coefficient, and the standard error. Some of the papers had 1000-genome reference data for specific populations[6][17].

Usually, quite a high number of genes are tested for correlation using metaCCA. This number is often in thousands which then further needs to be pruned to have a statistically significant set of genes. The p-value for each gene from metaCCA is used to filter the significant genes based on Bonferroni correction[6][22][17][18]. Since there are multiple hypothesis tests with quite a large set of genes, there might be increased chances of type-1 error; meaning there might be increased chance of rejecting significant genes[33]. Bonferroni correction scales the significance threshold value for each individual by dividing the statistical threshold value by the number of hypotheses or as in this case, the total number of tested genes[31].

Bonferroni correction aggressively reduces the significant genes. Even after reduction, there are further steps where the list of genes is reduced. The idea of such aggressive reduction is to not only have statistically significant genes but also have genes that can be used to understand the underlying biological functions using more conceptual analysis which helps make sense of such sets of genes.

2.4 Relationship between Blood Pressure, Pulse Rate and Cardiovascular disease

High blood pressure is strongly related to cardiovascular disease[13]. High blood pressure puts a strain on vascular system components which leads to problems as a person ages[13]. The known association of blood pressure directly affects the performance of not only the heart but blood flow in our body.

When it comes to pulse rate, a higher heart rate is associated with cardiovascular problems[36]. The risk of CVD observed with high blood pressure was comparable to risk due to high heart rate[36]. Pulse rate and blood pressure are related to each other.

2.5 Analysing Genes obtained from metaCCA

After we get genes from metaCCA, it is a good practise to analyse it further to find what the sets of genes do[22][6][17][18].

Often a gene-based analysis helps to understand the genes related to a trait. The aim is to find genes related to a trait and further analyse it. There are chances of a stronger association of certain genes to a trait than pleiotropic genes[6]. The genes from the gene-based analysis are often used to find pleiotropic genes and are used to refine the list of genes obtained from metaCCA[6][17].

Enrichment analysis helps to understand what certain genes do and what functions they take part in. There are two types of enrichment analysis that are used often[6][17][18][22]: pathway enrichment analysis and GO enrichment analysis.

Pathway enrichment analysis focuses to match the set of genes to certain pathway databases like wiki-pathway database[29]. These types of pathway databases have links to genes and what biological process they are involved in. Pathway enrichment analysis makes it easier to analyse a bunch of genes.

Gene Ontology, GO is an initiative to present the genes in 3 computational formats[1][1]. It shows the gene knowledge in biological processes, molecular functions and cellular components[1][1]. Using the GO terms, one can better understand how a gene affects the biology of our body.

Chapter 3

Execution

Context

This section focuses on the execution of the multivariate analysis of the chosen phenotypes, for cardiovascular diseases, using metaCCA. There are different sections exactly in order of how the analysis was executed. Each section elaborates thoroughly on all the important details required to interpret and reproduce the analysis.

3.1 GWAS Datasets

A major part of this analysis was to collect the right kind of GWAS summary datasets. This section focuses on the collection process of the data, the source of the data and elaborates on the original study as well.

3.1.1 Atlas GWAS catalog

Since GWAS summary statistics datasets are anonymous, it makes it easier to share between multiple parties. Sometimes, these datasets also can be posted online. The websites that host such datasets are often a good way to find the GWAS summary datasets of interest. One such source from where I found my GWAS datasets were from atlas website[49].

GWAS Atlas was an effort to document, present and host over 4000 publically available GWAS summary statistics datasets[49]. These GWAS studies were collected across many diseases and covered about 558 phenotypes/traits[49]. The effort was primarily to understand many complex diseases by exploring genetic architectures[49]. A good outcome of this study was the atlas website where one can use the collected GWAS summary datasets as well[49].

At the time of collecting the datasets for the original atlas research, the researchers also used UK BioBank data[49]. The collected UK BioBank data included about 600 traits[49]. The collected phenotypes were selected based on the individuals in the study with a lower limit of 50,000 for sample size and 10,000 for healthy individuals[49]. The atlas website also has analytics regarding these datasets which comes in handy while selecting these datasets.

3.1.2 UK BioBank

It is important to know the source of the study from where the GWAS summary statistic datasets are produced. UK BioBank GWAS summary datasets are very desirable for the aim of this thesis as the phenotypes they recorded had a large number of individuals[46]. With statistical analysis, a large number of individuals capture more genetic variations which can lead to discovering rare genetic associations with diseases.

Since the first GWAS studies in 2005[12], thousands of GWAS have been published covering a large number of phenotypes[49]. UK BioBank is a very large study with over 500,000 participants that covered thousands of traits[46]. The participants aged from 40 to 69 years[46]. This study was initially carried out in 2006-2010 but has also maintained and updated the data throughout the years[46]. UK BioBank was one of the few and rare initiatives that collected a large sample of data[46]. Also, since the focus of this thesis favors exploring the pleiotropic genes in the UK population, this biobank was a suitable

source. All the information regarding the UK BioBank can be accessed through the website (https://www.ukbiobank.ac.uk/)[46].

3.1.3 Datasets

The collected datasets from the atlas website have an atlas ID, unique for the datasets and also a UK biobank phenotype ID which can be used to access details regarding the study on the UK biobank website[46]. The following GWAS summary datasets were downloaded from the atlas website[49]:

1. Pulse Rate (Atlas ID: 3188, UKB Phenotype ID: 102):

2. Diastolic Blood Pressure (Atlas ID: 3379, UKB Phenotype ID: 4079):

3. Systolic Blood Pressure (Atlas ID: 3380, UKB Phenotype ID: 4080):

Phenotype	SNP Count	Population
Pulse Rate	10,534,620	European
Diastolic Blood Pressure	10,534,620	European
Systolic Blood Pressure	10,534,620	European

Table 3.1: Summary of GWAS datasets. Sourced from atlas website[49]

These datasets were part of the atlas study, collected in 2019[49]. All three phenotypes were measured using automated machines[46]. The sample size for the calculated summary statistics was 361,411[49]. Each of the dataset had 10,534,620 SNPs[49]. More details of the datasets are given in the table-3.1.

3.2 Pre-Processing

Pre-processing the data aims to prepare the data for analysis. Not only it filters the data, but it also makes sure the underlying assumptions of the data hold true. All of the files are in tab-delimited text files. The preliminary pre-processing steps were as follows:

1. Each of the datasets was checked for null/missing values. If found, the whole row was dropped.
2. All the data types of each column in each of the datasets were checked and any unusual or conflicting rows were removed.
3. Three main data variables in each of the datasets that were necessary for tracking the SNPs were:
   (a) Chromosome
   (b) Base-Pair Position
   (c) Alleles

   All these features were made sure to be of homogenous data types and any conflicting values were removed.
4. SNP-related features in the list-3 can be used to uniquely identify each row in the datasets. A unique key was used, composed of: [Chromosome: Base-Pair Position: Alphabetically sorted alleles] to check for any duplicate entries in the dataset. All the duplicates were dropped, keeping only the first instance of it.
5. The datasets also had RS ID columns. The SNP IDs in "RS" terms refer to a reference SNP ID used by researchers and other databases to uniquely identify each SNP. These "RS" terms will be required here to merge and compare different datasets based on SNPs.

   Any RS ID that didn't follow the usual format, where each ID name starts with "rs", was dropped. Any row with duplicate RS IDs was dropped as well.

Most of the above filters didn't change the size of the original data but were necessary filters to verify the datasets.

After preliminary pre-processing, the datasets were merged on a unique key as mentioned in point 4 in the above list. Any uncommon/missing IDs across the three datasets were dropped.

After merging the three datasets, the format of the merged file was according to the metaCCA implementation in R[7]. The merged file had the standard error and linear regression coefficient for each of the SNP along with alleles.

3.3 Calculating S_{YY}

Out of the three components of metaCCA, calculating the phenotype-phenotype correlation matrix is one of them[7]. After the preprocessing step, the merged file is used with a utility function from metaCCA implementation in R to calculate the S_{YY} matrix[7]. As discussed in chapter-2, the merged file has all the SNPs after pre-processing which are used to calculate this correlation matrix.

3.4 Pre-Processing for S_{XX} and S_{XY}

3.4.1 Reference Data for the Study Population

One of the prerequisites to calculating the S_{XX} matrix is to use a reference dataset of the study population[7]. The reference database needed for the datasets according to the atlas website should be of European ancestry, table-3.1[49]. A suitable reference database I used was the 1000-genome database for the European population[8], which is also the one used in the metaCCA paper[7].

1000-Genome Database

1000-genome database is a resource for creating a haplotype map[8]. Haplotype map tries to find a connection between genes or genetic variations and diseases[9]. The 1000-genome project resulted in a haplotype map of 1,092 individuals from 14 populations in its initial phase[8]. Using such a reference database helps compare only those regions of interest in DNA that are known to be associated with diseases, in a particular population.

3.4.2 Filtering of the Merged File

After the pre-processing step which resulted in a merged file, it is again filtered to have the right set of SNPs. It is often a good practice to remove low-quality SNPs from the dataset as it reduces the chances of bias in the results[34]. Out of some metrics for filtering SNPs, the only suitable metric available in the dataset was minor allele frequency (MAF).

Minor Allele Frequency

Minor allele frequency (MAF) is the frequency of the least common allele found at a specific position in the genome[27]. For low frequency of variations, it is often difficult to find associations for those variants[27].

While reading through some of the literature with application of metaCCA, different papers[6][17][22][18] had varying degree of MAF threshold to filter the data. I used one of the most common threshold values. I chose a MAF threshold of 0.01 (1%). It ensures that the SNPs that might not have much information, in terms of variations, are removed. Keeping a low threshold value doesn't diminish the possibility of finding new associations in the data. After filtering the merged file with the MAF criteria, about 32% of data was lost.

3.4.3 Pre-processing 1000-Genome Data

1000-genome reference data for European ancestry is required to be filtered for the SNPs of interest needed for the analysis. Due to limited computational resources, I used tools and software which could help me overcome the limitations and let me handle very large files. To handle the 1000-genome file which was upwards of more than 150GB, I used a software called Plink[37].

Plink

Plink is an open source command line tool created to efficiently handle, manipulate and pre-process GWAS datasets[37]. Written in low-level programming languages, makes it efficient and faster to use[37]. Being a popular choice[6][17][22][18], there is good compatibility of this tool with different types of data which makes it easier use.

Using plink, I extracted only the common SNPs between 1000-genome data and MAF filtered data (as per section-3.4.2). After merging, about 1% of the data was lost.

Reference Allele

It is important to understand the concept of alternative alleles and reference alleles before we proceed further. For the GWAS datasets, each position in the genome has an observed allele from a sample known as an alternative allele and a reference allele from a set of genes for that organism, known as a reference allele. The idea of reference allele is not to be precise with its sets of genes but rather to give all possible sets of genes for that organism. The reference genome build used for our data is GRCh37/hg19[49].

Recoding 1000-Genome Data

The plink files of 1000-genome data still needed to be recoded in terms of the total genotype counts (GT). After extracting the common SNPs between the 1000-genome dataset and MAF filtered dataset, I recoded the 1000-genome data as follows (sourced from the unpublished paper of my supervisor[47]):

1. if the alternative allele of the MAF filtered dataset equals the first letter of the 1000-genome alternative allele, then it is recoded as GT=1.

2. if the alternative allele of the MAF filtered dataset equals the second letter of the 1000-genome alternative allele, then it recoded as GT=2.

3. if the alternative allele of the MAF filtered dataset equals the third letter of the 1000-genome alternative allele, then it recoded as GT=3.

4. if the alternative allele of the MAF filtered dataset equals the reference allele of the 1000-genome data, then it recoded as GT=0.

After recoding the file consists of the genotype counts which will be used to calculate the genotype-genotype correlation matrix, S_{XX}.

3.5 Linkage Disequilibrium Pruning

3.5.1 Linkage Disequilibrium (LD)

There are many influencing factors that can structure a genome. Sometimes, It helps to understand this non-random association of alleles at certain genome positions[43]. This is known as Linkage Disequilibrium[43]. In a certain population, there may exist genetic forces which can influence the structure of a genome[43]. Linkage disequilibrium is one of the sensitive indicators to detect it[43].

Assuming the data is randomly shuffled, SNPs with a higher frequency of association of their alleles are said to be in linkage disequilibrium [27]. The concept of linkage disequilibrium is important because it directly affects the correlations between SNPs[27].

3.5.2 Pairwise Linkage Disequilibrium

LD helps to detect patterns between SNPs[41]. These patterns are often used in studies to identify SNPs that are affected by other SNPs[41]. A pairwise approach to the linkage disequilibrium solves this problem by comparing each SNP to other SNPs in pairs. Some approaches take an iterative approach where they compare variants in certain regions.

3.5.3 Pruning using Plink

Sometimes, the quality of the analysis of GWAS datasets can be affected by the high pairwise linkage disequilibrium[4]. Due to this, we have to prune the genotype data for high linkage disequilibrium[4]. Pruning based on pairwise linkage disequilibrium also reduces the computational cost and it aggressively reduces the data as well.

To prune the genotype data obtained after recoding, I used the plink tool[37]. Plink uses a repeated, window-based pruning algorithm, which uses the correlation metric r^2 threshold to prune the SNPs[37]. The algorithm filters based on the correlation coefficient between an SNP and other SNPs in a particular window[37]. The window here means the no of markers/SNPs or base pairs, often measured in kb where 1kb = 1000 base pairs.

For pruning my parameters were as follows:

1. Window Size: 1000000
2. Step Size: 5
3. r^2 threshold: 0.2

Increasing window size will take that many SNPs for LD calculation. Large window size helps in faster computation as fewer comparisons are made but it also means there might be regions of high LD.

To mitigate the chance of high LD after pruning, I used a smaller step size. The idea was that a larger correlation matrix will be computed but will use a small step size to thoroughly find high LD pairs while using maximum computational resources.

The r^2 is the final and main metric for pruning criteria. I used a lower threshold value to aggressively remove more pairs of high LD SNPs, the lower the r^2 threshold, the higher the pruning. Based on my research most papers used a 0.1 to 0.2 threshold for meta-analysis[6][17][22][18][47].

After pairwise linkage disequilibrium pruning, about 95% of the data was pruned.

3.6 Gene Annotation

Gene annotation is the process of annotating human genes to the data. A dataset of all known human genes is often used to track gene regions. In an idealistic situation, a genome build should include all genetic information of humans[19]. But often, due to diversity, it's difficult to capture such variations[19]. Human genome build is essentially used as reference data from the samples they received. From these samples, researchers try to create a reference map of what regions in a chromosome often have what type of gene. The reference data for our datasets were derived from GRCh37/hg19 human genome build[49]. Using this list of genes, we can filter our dataset for the specific types of genes.

For the gene annotation process, I used plink and GRCh-37/hg-19 human genome build. The criteria for annotation were as follows:

1. Only those SNPs which had an exact match for a position were selected.
2. Any type of mutation like missense, nonsense and frameshift were not included.
3. SNPs with only a single gene were selected.

Based on this criteria, using plink, at the end of this process about 15,198 genes were left out of 26,291 genes. There was a further reduction of data where about 57% of the data was left.

This was a critical step because the data now has the right set of markers/SNPs which matches with the reference data precisely. Now, the data also has a list of genes attached to it which will be used for the further process.

3.7 Applying metaCCA

After completing the above-mentioned steps, there were 166,130 SNPs left and 15,198 genes were left. Due to limited computational resources for calculating the correlation matrix, I tried to process the data in batches.

3.7.1 Calculating S_{XX}

After gene annotation, the 1000-genome data further needed to be calculated for S_{XX}. The correlation matrix of 166,130 SNPs would have resulted in a 166,130 X 166,130 matrix which required a much higher amount of computational memory resources than I had available to me. To overcome this problem, I filtered the data per chromosome and then calculated the correlation matrix. This reduced the computational time and also was easily calculated on my computer.

3.7.2 Calculating S_{XY}

After the gene annotation, the annotated SNPs were used to filter the MAF filtered dataset (as discussed in section-3.4.2) to create the ${}_SXY$ data-frame. The S_{XY} data-frame is formatted as required by the metaCCA implementation in R[7].

The final S_{XY} data-frame was filtered to remove any SNP with a standard error equal to 0, across the three traits. This is because if the standard error of the linear regression coefficient is 0, it means the model failed to any linear relationship. Hence, such instances are dropped from the S_{XY} data-frame.

3.7.3 Algorithm

metaCCA can work with the data of at least 1 chromosome[7]. The algorithm used to apply the metaCCA model addresses the computational resource limitation by processing the data in batches.

$GT \leftarrow$ genotype data for calculating S_{XX}
foreach *chromosome* **in** S_{XY} **do**
  $S_{XY}\_temp \leftarrow S_{XY}[\ chromosome\ in\ S_{XY} = chromosome\ ]$
  $GT \leftarrow GT[\ chromosome\ in\ GT = chromosome\ ]$
  $S_{XX} \leftarrow corr(GT)$
  foreach *gene* **in** S_{XY} **do**
    $S_{XY}\_temp \leftarrow S_{XY}\_temp[\ gene\ in\ S_{XY}\_temp = gene\ ]$
    $S_{XX} \leftarrow S_{XX}\_temp[\ gene\ in\ S_{XX}\_temp = gene\ ]$
    $results \leftarrow metaCca(S_{YY},\ S_{XY},\ S_{XX})$
    $save(results)$
  end
end

Algorithm 3.1: Algorithm used to apply metaCCA

The algorithm-3.1 which applies the metaCCA model to find the pleiotropic genes, works as follows:

1. First, the program loops over each chromosome in the S_{XY}.
2. For the chromosome, S_{XY} is further filtered to only have SNPs in that chromosome.
3. The genotype data is also filtered for SNPs in each chromosome.
4. S_{XX} is computed from the filtered genotype data
5. The program again loops over each gene in the filtered S_{XY} having only a single chromosome data.
6. S_{XY} and S_{XX} are again filtered for each gene
7. The metaCCA algorithm is applied and the results are recorded.

While calculating the genotype correlation matrix S_{XX}, the SNPs with zero correlation were dropped. Filtering these SNPs which were not correlated at all further reduces the chances of getting incorrect results.

The metaCCA model required the coefficients of the S_{XY} matrix to be standardised[7]. While applying the model, this was handled in the parameters given to the metaCCA function in R. It's also worth mentioning that for instances where only one SNP was found for each gene in a chromosome, the univariate version of the metaCCA was used[7].

3.8 Bonferroni Correction

As discussed in chapter-2, Bonferroni correction is essential for filtering the statistically significant genes. The total number of tested genes were 15,198. The p-value threshold for filtering was 0.05/15,198. There were some instances of p-values with the value 0. This was because any number less than 10^{-300} in R is converted to 0. I imputed these 0 p-values with 10^{-300} value.

3.9 Gene-Set Analysis and Enrichment Analysis

For gene set analysis, a model named MAGMA was used[11]. MAGMA uses multiple regression models to analyse the genes[11]. The use of the regression model helps lower the risk of biased results due to statistical power[11]. Using MAGMA, I analyzed the overlapping gene count between traits. The results of MAGMA were obtained from the atlas website[49].

For Enrichment analysis, enrichr web tool was used[5][20][51]. The online tool analysed the set of 860 genes for both pathway enrichment analysis and GO enrichment analysis. Wiki-pathway 2021 database is used for pathway analysis.

Chapter 4

Results

Context

This chapter focuses on discussing the results of the analysis. Each section also talks about results obtained from different steps.

4.1 S_{YY} Matrix

The phenotype correlation matrix calculates the correlation between all the traits. From the figure, it is evident that the systolic and diastolic blood pressure traits are correlated. This means that when systolic blood pressure increases diastolic blood pressure can be expected to increase as well. This result is expected as it follows the assumption of the linear relationship between the two[40]. Pulse rate has a very low correlation with systolic blood pressure compared to diastolic blood pressure.

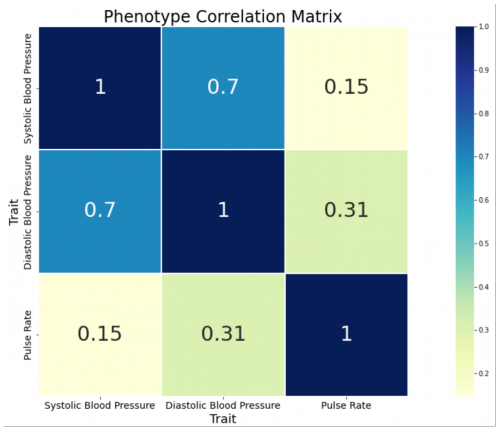

Figure 4.1: Phenotype-Phenotype Correlation Matrix

4.2 metaCCA Genes

After applying the metaCCA model and filtering genes using Bonferroni correction, I got about 860 statistically significant genes. Even though statistically significant, these genes have to be further analyzed to make any meaningful conclusions.

The manhattan plot summarises the total tested genes. It shows the significant genes at each site in the chromosome. The cut-off line is the Bonferroni adjusted p-value threshold which was 3.29×10^{-6}. The most significant genes were present at the 1^{st} chromosome which had 86 genes in it.

The canonical correlation barplot shows the top 100 genes and their correlation values. The correlation values, in all the significant 860 genes, varied from 0.007981 to 0.815643.

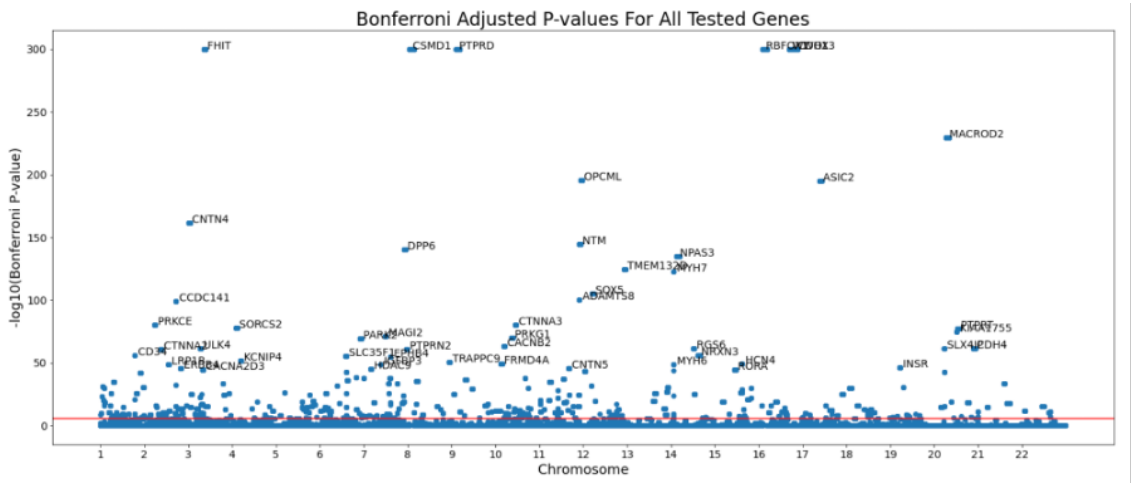

Figure 4.2: Manhattan Plot of all tested genes

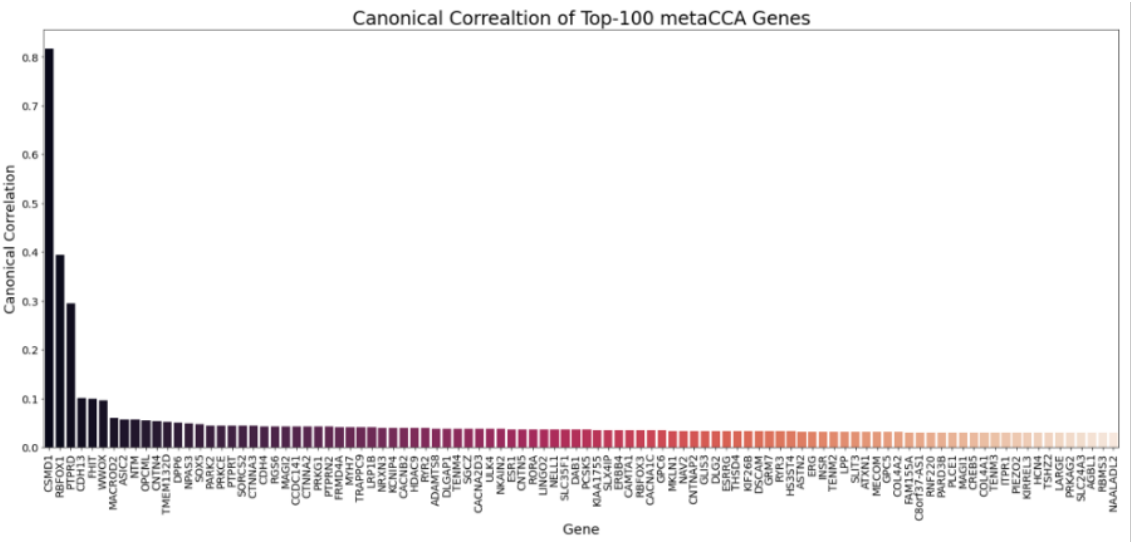

Figure 4.3: Canonical Correlation of top 100 genes

There are 6 genes that have very low p-values and have the highest correlation value. These genes are FHIT, CSMD1, PTPRD, RBFOX1, WWOX, and CDH13. Even though these are the most statistically significant genes and also have high canonical correlation values, they still might not give any meaningful information as the context of what these genes are, is missing.

Diseases associated with these genes are:

1. FHIT: This is a protein-coding gene[44]. Cancer diseases like Renal Cell Carcinoma, Nonpapillary and Sporadic Breast Cancer are associated with this gene[44][38].
2. CSMD1: This is a protein-coding gene[44]. This gene is associated with diseases like Schizophrenia[44][38].
3. PTPRD: This too is a protein-coding gene[44]. Restless Legs Syndrome and Chromosome 9P Deletion Syndrome are some of the diseases that are associated with this gene[44][38].
4. RBFOX1: This is a protein-coding gene as well[44]. Benign Epilepsy With Centrotemporal Spikes and Colorectal Cancer are some of the diseases associated with this gene[44][38].
5. WWOX: This is a protein-coding gene[44]. Spinocerebellar Ataxia, Autosomal Recessive 12 and Developmental And Epileptic Encephalopathy 28 are some of the diseases associated with this gene[44][38].
6. CDH13: This is a protein-coding gene[44]. Vacterl Association and Seminoma are some of the diseases that this gene is associated with[44][38].

The top 100 genes sorted based on p-values and correlation values are shown in the barplot. Most of the genes are concentrated within the correlation value of 0.1 to 0.

4.3 Gene-Set Analysis

Using MAGMA, each trait was tested for genes associated with it. The heatmap plot shows the overlap of genes with other genes associated with each trait. The highest overlap of SNPs is observed for the systolic and diastolic GWAS data. Systolic trait had the most combined overlapped genes with other traits, at about 786 genes.

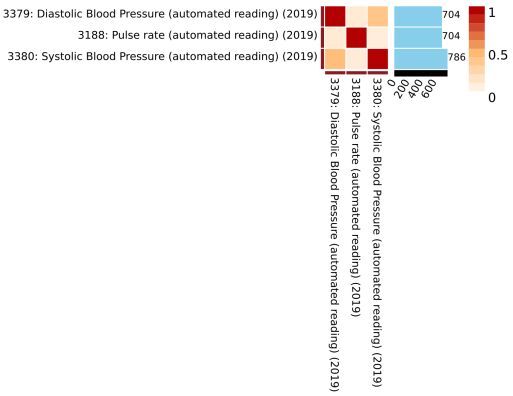

Figure 4.4: Overlapped genes between traits. Sourced from atlas website[49]

4.4 Pathway Enrichment Analysis

From the results of pathway enrichment analysis, significant enrichment is observed in 3 pathways in the Wiki-pathway database.

These pathways in order of their significance are:

1. Calcium Regulation in the Cardiac Cell WP536[29]: This was the most significant pathway which is directly related to cardiovascular system diseases[29]. This pathway describes the calcium signaling in the heart[28][29]. This signaling is known to activate important functions like cardiac muscle contraction[28].

2. Pathways Regulating Hippo Signaling WP4540[29]: This pathway is related to an aggressive type of cancer known as malignant mesothelioma[29][42]. This pathway affects many other pathways and factors[29][42].

3. Arrhythmogenic Right Ventricular Cardiomyopathy WP2118[29]: Arrhythmogenic right ventricular cardiomyopathy is a rare form of cardiomyopathy (a heart disease related to the pumping of the blood)[29][16]. This disease often leads to sudden death, and heart failure[29].

The top-5 pathways and overlapped genes are as follows:

1. Calcium Regulation in the Cardiac Cell WP536 - CHRM2, RYR1, RYR2, ITPR1, CACNA1A, ITPR2, CACNA1D, CACNA1C, RYR3, GNGT1, GJA1, GNG7, RGS6, RGS7, KCNJ5, PRKCB, PRKCE, ATP1B3, PRKCA, ATP2B1, ATP1B1, GNAO1, PKIB, ADCY9, GNAS, GNB4

2. Pathways Regulating Hippo Signaling WP4540 - TCF7L2, TCF7L1, SMAD3, PRKCB, PRKCE, INSR, MST1, PRKAG2, PRKCA, EGFR, IGF1R, CDH4, LATS2, CDH2, PRKD3, CDH11, GNAS, CDH13, PLCB1, MET

3. Arrhythmogenic Right Ventricular Cardiomyopathy WP2118 - DSP, RYR2, TCF7L2, TCF7L1, ITGB5, CACNA2D1, CACNA2D3, CACNA1D, ACTN4, CACNA1C, CACNB2, GJA1, CDH2, PKP2, CTNNA3, CTNNA2, ITGA9

4. Myometrial relaxation and contraction pathways WP289 - RYR1, RYR2, PRKCB, NOS3, IGFBP3, PRKCE, PDE4D, ITPR1, ITPR2, PRKCA, CRHR1, RYR3, GNGT1, PKIB, GJA1, ADCY9, GNG7, PLCG2, GNAS, GNB4, RGS6, RGS7

5. Resistin as a regulator of inflammation WP4481 - AKT2, AKT3, ITPR1, PLCG2, PLCE1, MAPK1, PLCB1, PLCD3, RELA

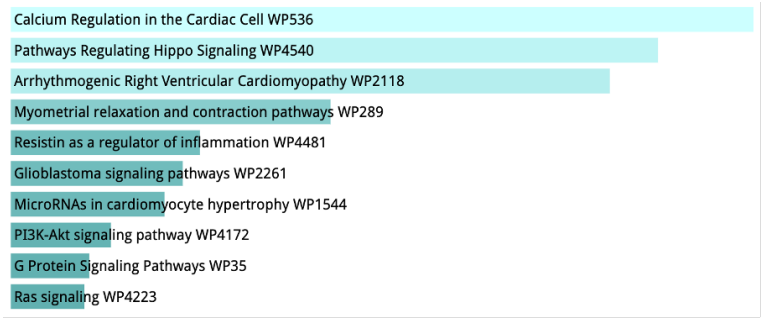

Figure 4.5: Top 10 Pathways in Wiki-Pathway Database 2021. Sourced from Enrichr[51][20][5]

4.5 GO Enrichment Analysis

The significant GO terms for biological processes are related to heart functions. These terms included heart contractions which match the functions of the measured traits.

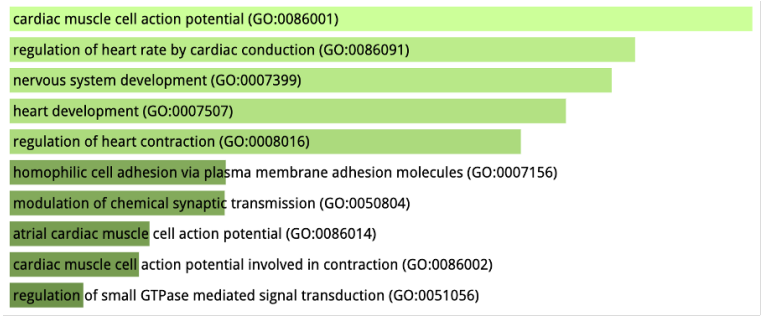

Figure 4.6: Top-10 biological processes GO terms. Sourced from Enrichr[51][20][5]

The top-5 GO biological processes and overlapped genes are as follows:

1. cardiac muscle cell action potential (GO:0086001) - HCN4, SCN10A, KCND3, CACNA2D1, KCNQ1, PKP2, CACNA1D, ANK2, CACNA1C, SCN5A, ATP1B1, FGF12

2. regulation of heart rate by cardiac conduction (GO:0086091) - HCN4, DSP, KCNJ5, KCND3, CACNA2D1, CACNA1D, ANK2, CACNA1C, CACNB2, KCNQ1, PKP2, CTNNA3, SCN5A

3. nervous system development (GO:0007399) - CHRM2, HDAC4, YAP1, NRXN1, WDR62, SPG7, TRAK1, KALRN, NPAS2, FGF5, ATXN1, RELN, EFNB3, CDH2, BTBD1, ERBB4, DNER, DPF3, EP300, SOX6, ASIC2, KIRREL3, FARP1, RBFOX1, MBNL1, MEF2C, SZT2, JAG1, RBFOX3, DSCAM, CNTN6, ZBTB16, FN1, FBXL17, CRIM1, NRG1, NAV2, NGF, ARHGAP26, DNM3, NELL1, SDK1, DLG4, CNTN4, MET, FGF12, SDK2, SHANK2

4. heart development (GO:0007507) - BMPR2, GATA4, CACNA1C, PCSK5, HDAC9, ERBB4, PLCE1, MEF2D, EPHB4, MYBPC3, MEF2C, MKKS, INSR, FN1, TSC2, SGCZ, TGFBR3, DLC1, PKP2, RBM20, ZFPM2, ZFPM1, FGF12, MYH6, FBN1, MYH7

5. regulation of heart contraction (GO:0008016) - RYR1, HCN4, RYR2, NPR1, CELF2, PDE4D, TPM1, ITPR1, ITPR2, ATP1B3, ANK2, ATP2B1, ATP1B1, RYR3, AGT, TRDN, SCN10A, KCNQ1, MYH6

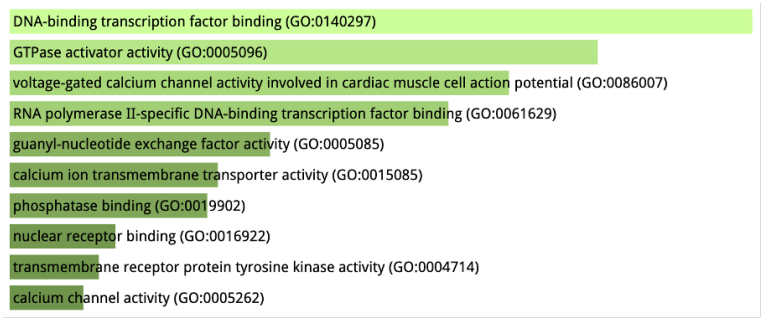

Figure 4.7: Top-10 GO terms related to molecular functions. Sourced from Enrichr[51][20][5]

The top-5 GO molecular functions and overlapped genes are as follows:

1. DNA-binding transcription factor binding (GO:0140297) - YAP1, HDAC4, SPI1, GATA4, CXXC5, HDAC9, RELA, PRDM16, EP300, MEF2D, GTF2I, TCF7L2, MEF2C, CREBBP, MKKS, SMAD3, PRRX1, NFATC1, WWP2, RBL2, DGKQ, MAP3K10, ZFPM2, TP53, ZFPM1
2. GTPase activator activity (GO:0005096) - DOCK4, DOCK8, ITSN1, RASGRF1, ASAP2, ARHGAP15, KALRN, PREX1, ABR, FGD4, FGD5, ARHGAP42, PLCE1, SBF2, RGS6, RGS7, VAV3, FARP1, EIF2B4, TSC2, ARHGEF18, MYO9B, ARHGAP27, ARHGAP26, ARHGAP24, ARFGAP2, ARHGAP10, TBC1D1, DLC1, RIN3, RGL1, PLCB1, DOCK1
3. voltage-gated calcium channel activity involved in cardiac muscle cell action potential (GO:0086007) - CACNB2, CACNA2D1, CACNA1D, CACNA1C
4. RNA polymerase II-specific DNA-binding transcription factor binding (GO:0061629) - PTPRT, HDAC4, BCAS3, TCF7L2, CREBBP, MKKS, SPI1, SMAD3, PRRX1, GATA4, NFATC1, WWP2, ACTN4, RELA, NCOR2, RBL2, EP300, TACC2, ZFPM2, TP53, ZFPM1, GTF2I
5. guanyl-nucleotide exchange factor activity (GO:0005085) - VAV3, EIF2B4, FARP1, DOCK4, DOCK8, ITSN1, RASGRF1, ARHGEF18, KALRN, PREX1, ABR, FGD4, FGD5, PLCE1, RIN3, RGL1, SBF2, DOCK1

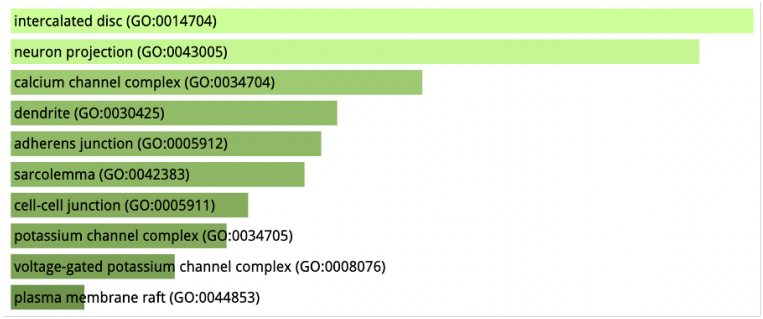

Figure 4.8: Top-10 GO terms related to cellular components. Sourced from Enrichr[51][20][5]

The cellular GO components are also related to the heart. The most significant molecular function, intercalated disc (GO:0014704)[5][20], is part of a cardiac muscle which helps in cardiac muscle contraction[48]. Most of the significant components were related to heart functions. The top-5 GO molecular functions and overlapped genes are as follows:

1. intercalated disc (GO:0014704) - DSP, GJA1, CDH2, NRAP, PKP2, CTNNA3, ANK2, SCN5A, ANK3, ATP1B1
2. neuron projection (GO:0043005) - CHRM2, CNTNAP2, APP, TENM2, TENM3, TENM4, RASGRF1, HTR4, IGF1R, SCGN, CDH2, GRM7, DNER, ANKS1A, EPHB4, KPNA1, KIRREL3, CHRNB1, PACRG, DSCAM, MAGI2, ANK2, ANK3, NGF, DNM3, SCN10A, PALLD, MAP1B, CDH13, RIN3, MAPT, SHANK2, SPTBN4, TRAK1, CRHR1, INPP5A, RELN, CTNNA2, CCDC141, MARK3, FARP1, HOMER2, WFS1, CNTN6, PLK2, INSR, DLG2, DLG4, STRN, CNTN4

3. calcium channel complex (GO:0034704) - RYR1, RYR2, CACNB2, TRPC4, CACNA2D1, PDE4D, CACNA2D3, CACNA1D, CACNA1C, RYR3

4. dendrite (GO:0030425) - CHRM2, CNTNAP2, TENM2, HTR4, TRAK1, SCGN, INPP5A, RELN, GRM7, DNER, MARK3, KPNA1, KIRREL3, FARP1, DSCAM, HOMER2, WFS1, INSR, PLK2, MAGI2, ANK3, NGF, DLG4, MAP1B, RIN3, STRN, MAPT

5. adherens junction (GO:0005912) - MAGI1, SPTBN4, JAG1, BMPR2, CTNND2, PTPRM, FRMD4A, PARD3B, ARVCF, CDH2, CCDC85A, CDH11, PKP2, FAT2, CTNNA3, CTNNA2, FERMT2

Chapter 5

Discussion

Context

This chapter focuses on summarising the results and talks about critical stages of implementation. It also focuses on any point of failure that might change the results.

5.1 Genes from metaCCA

The gene set obtained from metaCCA was quite big to analyze for each gene. This set also included many genes associated with different kinds of diseases including cardiovascular diseases. The top genes obtained from the set of the metaCCA genes are FHIT, CSMD1, PTPRD, RBFOX1, WWOX, and CDH13. Even though they are the most correlated and statistically significant genes out of all the tested genes, there are lacking the context of what they do. Some of these six genes are related to cancer diseases.

5.2 Gene-set Analysis

Gene set analysis using MAGMA was to look into overlapping significant sets of genes in the three traits. The analysis helped in understanding the pairs of traits which had the most overlapped genes. Confirming the results of S_{YY} which indicated a linear relationship between systolic and diastolic blood pressure, the most common set of genes was found between the systolic and diastolic blood pressure traits. A good indicator of this analysis was the high amount of total overlap of genes. The lowest count of overlapped genes is 704 and the highest is 786.

A better way of analysing these genes is by looking for the association of subsets of genes to any critical biological process or components. Hence, enrichment analysis helps to understand this very function[6].

5.3 Enrichment Analysis

To find subsets of genes in the genes obtained from the metaCCA stage, a sound strategy is to use enrichment analysis[6][17][22][18]. Two types of enrichment analysis, pathway analysis and GO term analysis are used to find these subsets of genes related to biological components and processes.

Pathway Analysis

The pathway analysis, conforming to the wiki-pathway 2021 database[29], resulted in the 3 most significant pathways. Since the sets of metaCCA genes had significant genes related to cancer diseases, the "Pathways Regulating Hippo Signaling WP4540" pathway was found to be significant as well. Interestingly, this pathway is related to an aggressive form of cancer. Apart from this pathway, the other two pathways confirm the association of a subset of genes with heart-related components.

GO Analysis

GO terms help understand the association of subsets of genes to critical biological processes, functions and components. The overlapped genes associated with GO terms relating to the three biological components

do indicate association with critical cardiac processes and biological components. There were a few terms not associated with cardiac processes.

5.4 Critical Evaluation

Even though the analysis was elaborate, covering various aspects, it is fair to critic on some points where a potential failure or improvement can be made. Some of these points are as follows:

1. Linkage Disequilibrium, used to prune SNPs, affects the correlation between SNPs. It is good to remove markers with high LD. Deciding the right values of window size, step size and r^2 threshold for the algorithm is critical to clean the dataset. My strategy was to take a larger window size with very small step size and a low r^2. Incorrect window size and step size combination might still leave some high LD regions. These could lead to biased results.

2. A good way of analysing the large set of genes obtained from the metaCCA model would have been to take the common genes from a gene-based analysis tool and the gene set obtained from the metaCCA[6][17]. I planned on using a tool to analyse genes per trait called Vegas-2[32]. In the span of 3 months of my thesis, this tool was not available on its site. This limited my results when I planned to refine my metaCCA set of genes further than I already did.

3. Context and biological background can improve such analysis.

Chapter 6

Conclusion

Context

This section aims to summarise the whole thesis while highlighting key decisions taken during the analysis and the results. It also includes a future work section that focuses on further improvements or experiments that I could have done given more time.

6.1 Summary

Diseases often cause changes or are caused/altered by changes in our bodies. With advancements in technology, we can now analyse and conduct studies that focus on such variations. These genome-wide association studies help understand these variations and help in focusing on the associations between these variations and diseases. Such variations often occur inside genes. Genes take part in biological processes and components.

Cardiovascular disease is a major cause of death worldwide[14]. This thesis was an effort to understand the association of genes for blood pressure and pulse rate phenotypes with cardiovascular disease. GWAS summary datasets are anonymous which makes it difficult to use them for canonical correlation analysis. metaCCA fills in this gap[7].

There were 4 stages that were followed for the analysis:

1. Data Pre-Processing: This stage involved pre-processing the data. The data was pre-processed based on many criteria like linkage disequilibrium, minor allele frequency and gene annotation. These steps were taken to reduce the bias in results.
2. Applying metaCCA: After the pre-processing stage, the metaCCA model was applied. The results from metaCCA were filtered for statistical significance based on Bonferroni-adjusted p-values.
3. Gene-set Analysis: The GWAS data were analyzed for genes associated with each trait and then checked for overlap between other sets of genes of other traits.
4. Enrichment Analysis: To analyse the subsets of genes, obtained from metaCCA results, and to understand what they do, enrichment analysis was done. This analysis included further two types of analysis:
   (a) Pathway Enrichment Analysis: This analysis was to find the gene pathways and correlate the results to what functions or diseases those genes are associated with. Wiki-pathway 2021 database was used[29].
   (b) GO Enrichment Analysis: The GO enrichment analysis tries to make sense of the gene set by understanding its biological process, molecular function and cellular component[1][2].

6.2 Project Status

The key points regarding the status of the project are as follows:

1. The results concluded that I found genes associated with cardiac processes which are further related to cardiovascular disease. Those genes are reported in chapter-4.

2. There was a good overlap between the genes I found and cardiac functions, but there were some associations with cancer as well. For some highly correlated genes, most of them were associated with cancer diseases. The second most significant pathway was directly associated with an aggressive form of cancer. These results are reported in chapter-4.

3. I planned on using the vegas-2[32] tool to analyse the genes associated with each trait. Taking the common genes from vegas-2 results and metaCCA results would have resulted in a refined gene list which I could have used to properly conclude my thesis. The vegas-2 website was having problems that restricted me from using it in any way possible.

6.3 Future Work

This topic can easily be extended or even can be more elaborate. Some of these points are as follows:

1. Linkage disequilibrium is a topic where deciding parameter values for the pruning algorithm can decide on the quality of the data. A more detailed analysis of LD can really give more insights into the data.

2. Refining metaCCA results can certainly help a lot for analysis. A technique of refinement is to use genes associated with traits and metaCCA genes to find the overlapping genes. These overlapped genes have a higher chance of giving more precise results and can be analysed in detail. This approach can be used to get a small, precise set of genes.

3. There were many stages where handling data and manipulating or computing it was difficult. Designing efficient algorithms using distributed computing or batch processing can certainly help speed up processes.

4. Having conceptual knowledge of biology or the bioinformatics side of things would have given better insights and analytical points. Taking a

Bibliography

[1] The gene ontology resource: enriching a gold mine. *Nucleic acids research*, 49(D1):D325–D334, 2021.

[2] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

[3] Prachi Bhatnagar, Kremlin Wickramasinghe, Julianne Williams, Mike Rayner, and Nick Townsend. The epidemiology of cardiovascular disease in the uk 2014. *Heart*, 101(15):1182–1189, 2015.

[4] Mario PL Calus and Jérémie Vandenplas. Snprune: an efficient algorithm to prune large snp array and sequence datasets based on high linkage disequilibrium. *Genetics Selection Evolution*, 50(1):1–11, 2018.

[5] Edward Y Chen, Christopher M Tan, Yan Kou, Qiaonan Duan, Zichen Wang, Gabriela Vaz Meirelles, Neil R Clark, and Avi Ma'ayan. Enrichr: interactive and collaborative html5 gene list enrichment analysis tool. *BMC bioinformatics*, 14(1):1–14, 2013.

[6] Yuan-Cheng Chen, Chao Xu, Ji-Gang Zhang, Chun-Ping Zeng, Xia-Fang Wang, Rou Zhou, Xu Lin, Zeng-Xin Ao, Jun-Min Lu, Jie Shen, and Hong-Wen Deng. Multivariate analysis of genomics data to identify potential pleiotropic genes for type 2 diabetes, obesity and dyslipidemia using meta-cca and gene-based approach. *PLOS ONE*, 13(8):1–16, 08 2018.

[7] Anna Cichonska, Juho Rousu, Pekka Marttinen, Antti J. Kangas, Pasi Soininen, Terho Lehtimäki, Olli T. Raitakari, Marjo-Riitta Järvelin, Veikko Salomaa, Mika Ala-Korpela, Samuli Ripatti, and Matti Pirinen. metaCCA: summary statistics-based multivariate meta-analysis of genome-wide association studies using canonical correlation analysis. *Bioinformatics*, 32(13):1981–1989, 02 2016.

[8] 1000 Genomes Project Consortium et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56, 2012.

[9] International HapMap Consortium et al. A haplotype map of the human genome. *Nature*, 437(7063):1299, 2005.

[10] Francis H Crick. On protein synthesis. In *Symp Soc Exp Biol*, volume 12, page 8, 1958.

[11] Christiaan A. de Leeuw, Joris M. Mooij, Tom Heskes, and Danielle Posthuma. Magma: Generalized gene-set analysis of gwas data. *PLOS Computational Biology*, 11(4):1–19, 04 2015.

[12] Albert O Edwards, Robert Ritter III, Kenneth J Abel, Alisa Manning, Carolien Panhuysen, and Lindsay A Farrer. Complement factor h polymorphism and age-related macular degeneration. *Science*, 308(5720):421–424, 2005.

[13] Flávio D Fuchs and Paul K Whelton. High blood pressure and cardiovascular disease. *Hypertension*, 75(2):285–292, 2020.

[14] Thomas Gaziano, K Srinath Reddy, Fred Paccaud, Sue Horton, and Vivek Chaturvedi. Cardiovascular disease. *Disease Control Priorities in Developing Countries. 2nd edition*, 2006.

[15] Wolfgang Karl Härdle and Léopold Simar. *Canonical Correlation Analysis*, pages 443–454. Springer Berlin Heidelberg, Berlin, Heidelberg, 2015.

[16] Thomas Herren, Philipp A Gerber, and Firat Duru. Arrhythmogenic right ventricular cardiomyopathy/dysplasia: a not so rare "disease of the desmosome" with multiple clinical presentations. *Clinical Research in Cardiology*, 98(3):141–158, 2009.

[17] XiaoCan Jia, YongLi Yang, YuanCheng Chen, ZhiWei Cheng, Yuhui Du, Zhenhua Xia, Weiping Zhang, Chao Xu, Qiang Zhang, Xin Xia, HongWen Deng, and XueZhong Shi. Multivariate analysis of genome-wide data to identify potential pleiotropic genes for five major psychiatric disorders using metacca. *Journal of Affective Disorders*, 242:234–243, 2019.

[18] XiaoCan Jia, YongLi Yang, YuanCheng Chen, Zhenhua Xia, Weiping Zhang, Yu Feng, Yifan Li, Jiebing Tan, Chao Xu, Qiang Zhang, Hongwen Deng, and XueZhong Shi. Multivariate analysis of genome-wide data to identify potential pleiotropic genes for type 2 diabetes, obesity and coronary artery disease using metacca. *International Journal of Cardiology*, 283:144–150, 2019.

[19] Chakravarthi Kanduri, Diana Domanska, Eivind Hovig, and Geir Kjetil Sandve. Genome build information is an essential part of genomic track files. *Genome biology*, 18(1):1–5, 2017.

[20] Maxim V Kuleshov, Matthew R Jones, Andrew D Rouillard, Nicolas F Fernandez, Qiaonan Duan, Zichen Wang, Simon Koplev, Sherry L Jenkins, Kathleen M Jagodnik, Alexander Lachmann, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids research*, 44(W1):W90–W97, 2016.

[21] Eleanor Lawrence. *Henderson's dictionary of biology*. Pearson education, 2005.

[22] Huanqiang Li, Ziling Mai, Sijia Yu, Bo Wang, Wenguang Lai, Guanzhong Chen, Chunyun Zhou, Jin Liu, Yongquan Yang, Shiqun Chen, et al. Exploring the pleiotropic genes and therapeutic targets associated with heart failure and chronic kidney disease by integrating metacca and sglt2 inhibitors' target prediction. *BioMed research international*, 2021, 2021.

[23] R Luengo-Fernández, J Leal, A Gray, S Petersen, and M Rayner. Cost of cardiovascular diseases in the united kingdom. *Heart*, 92(10):1384–1389, 2006.

[24] Jacqueline A.L. MacArthur, Annalisa Buniello, Laura W. Harris, James Hayhurst, Aoife McMahon, Elliot Sollis, Maria Cerezo, Peggy Hall, Elizabeth Lewis, Patricia L. Whetzel, Orli G. Bahcall, Inês Barroso, Robert J. Carroll, Michael Inouye, Teri A. Manolio, Stephen S. Rich, Lucia A. Hindorff, Ken Wiley, and Helen Parkinson. Workshop proceedings: Gwas summary statistics standards and sharing. *Cell Genomics*, 1(1):100004, 2021.

[25] Anubha Mahajan, Min Jin Go, Weihua Zhang, Jennifer E Below, Kyle J Gaulton, Teresa Ferreira, Momoko Horikoshi, Andrew D Johnson, Maggie CY Ng, Inga Prokopenko, et al. Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nature genetics*, 46(3):234–244, 2014.

[26] Teri A. Manolio. Genomewide association studies and assessment of the risk of disease. *New England Journal of Medicine*, 363(2):166–176, 2010. PMID: 20647212.

[27] Andries T Marees, Hilde de Kluiver, Sven Stringer, Florence Vorspan, Emmanuel Curis, Cynthia Marie-Claire, and Eske M Derks. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *International journal of methods in psychiatric research*, 27(2):e1608, 2018.

[28] Andrew R Marks et al. Calcium and the heart: a question of life and death. *The Journal of clinical investigation*, 111(5):597–600, 2003.

[29] Marvin Martens, Ammar Ammar, Anders Riutta, Andra Waagmeester, Denise N Slenter, Kristina Hanspers, Ryan A. Miller, Daniela Digles, Elisson N Lopes, Friederike Ehrhart, Lauren J Dupuis, Laurent A Winckers, Susan L Coort, Egon L Willighagen, Chris T Evelo, Alexander R Pico, and Martina Kutmon. WikiPathways: connecting communities. *Nucleic Acids Research*, 49(D1):D613–D621, 11 2020.

[30] Atul Mehta, Michael Beck, and Gere Sunder-Plassmann, editors. *Fabry Disease: Perspectives from 5 Years of FOS*. Oxford PharmaGenesis, 2006.

[31] R.G.J. Miller. *Simultaneous Statistical Inference*. Springer Series in Statistics. Springer New York, 2012.

[32] Aniket Mishra and Stuart Macgregor. Vegas2: Software for more flexible gene-based testing. *Twin Research and Human Genetics*, 18(1):86–91, 2015.

[33] Ron C Mittelhammer, George G Judge, and Douglas J Miller. *Econometric foundations pack with CD-ROM*. Cambridge University Press, 2000.

[34] Zahra Mortezaei and Mahmood Tavallaei. Novel directions in data pre-processing and genome-wide association study (gwas) methodologies to overcome ongoing challenges. *Informatics in Medicine Unlocked*, 24:100586, 2021.

[35] V Orgogozo, Alexandre E. Peluffo, and Baptiste Morizot. The " Mendelian Gene " and the " Molecular Gene " : Two Relevant Concepts of Genetic Units. In Virginie Orgogozo, editor, *Genes and Evolution*, volume 119 of *Current Topics in Developmental Biology*, pages 1–26. Elsevier, 2016.

[36] Christine Perret-Guillaume, Laure Joly, and Athanase Benetos. Heart rate as a risk factor for cardiovascular disease. *Progress in Cardiovascular Diseases*, 52(1):6–10, 2009. Heart Rate and Cardiovascular Disease.

[37] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A.R. Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I.W. de Bakker, Mark J. Daly, and Pak C. Sham. Plink: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575, 2007.

[38] Noa Rappaport, Michal Twik, Inbar Plaschkes, Ron Nudel, Tsippi Iny Stein, Jacob Levitt, Moran Gershoni, C. Paul Morrey, Marilyn Safran, and Doron Lancet. MalaCards: an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search. *Nucleic Acids Research*, 45(D1):D877–D887, 11 2016.

[39] Richard H Scheuermann, Werner Ceusters, and Barry Smith. Toward an ontological treatment of disease and diagnosis. *Summit on translational bioinformatics*, 2009:116, 2009.

[40] Giuseppe Schillaci and Giacomo Pucci. The dynamic relationship between systolic and diastolic blood pressure: yet another marker of vascular aging? *Hypertension research*, 33(7):659–661, 2010.

[41] Steven J Schrodi, Veronica E Garcia, Charley Rowland, and Hywel B Jones. Pairwise linkage disequilibrium under disease models. *European journal of human genetics*, 15(2):212–220, 2007.

[42] Yoshitaka Sekido. Targeting the hippo pathway is a new potential therapeutic modality for malignant mesothelioma. *Cancers*, 10(4):90, 2018.

[43] Montgomery Slatkin. Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, 9(6):477–485, 2008.

[44] Gil Stelzer, Naomi Rosen, Inbar Plaschkes, Shahar Zimmerman, Michal Twik, Simon Fishilevich, Tsippi Iny Stein, Ron Nudel, Iris Lieder, Yaron Mazor, et al. The genecards suite: from gene data mining to disease genome sequence analyses. *Current protocols in bioinformatics*, 54(1):1–30, 2016.

[45] Matthew Stephens. Correction: A unified framework for association analysis with multiple related phenotypes. *Plos one*, 14(3):e0213951, 2019.

[46] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3):e1001779, 2015.

[47] Amy Osborne Sumayah Binhazzaa and Colin Campbell. C-cass: a fast method for canonical correlation analysis based on summary statistics from genome-wide association studies.

[48] Yan Sun, Seung-Min Lee, Bon-Jin Ku, and Myung-Jin Moon. Fine structure of the intercalated disc and cardiac junctions in the black widow spider latrodectus mactans. *Applied Microscopy*, 50(1):1–9, 2020.

[49] Kyoko Watanabe, Sven Stringer, Oleksandr Frei, Maša Umićević Mirkov, Christiaan de Leeuw, Tinca JC Polderman, Sophie van der Sluis, Ole A Andreassen, Benjamin M Neale, and Danielle Posthuma. A global overview of pleiotropy and genetic architecture in complex traits. *Nature genetics*, 51(9):1339–1348, 2019.

[50] Bush WS and Moore JH. Chapter 11: Genome-wide association studies. *PLoS Computational Biology*, 2012.

[51] Zhuorui Xie, Allison Bailey, Maxim V Kuleshov, Daniel JB Clarke, John E Evangelista, Sherry L Jenkins, Alexander Lachmann, Megan L Wojciechowicz, Eryk Kropiwnicki, Kathleen M Jagodnik, et al. Gene set knowledge discovery with enrichr. *Current protocols*, 1(3):e90, 2021.

[52] Jian Yang, Teresa Ferreira, Andrew P Morris, Sarah E Medland, Pamela AF Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, Michael N Weedon, Ruth J Loos, et al. Conditional and joint multiple-snp analysis of gwas summary statistics identifies additional variants influencing complex traits. *Nature genetics*, 44(4):369–375, 2012.

[53] Zhi Yu, Josef Coresh, Guanghao Qi, Morgan Grams, Eric Boerwinkle, Harold Snieder, Alexander Teumer, Cristian Pattaro, Anna Köttgen, Nilanjan Chatterjee, and Adrienne Tin. A bidirectional mendelian randomization study supports causal effects of kidney function on blood pressure. *Kidney International*, 98(3):708–716, 2020.