Master Degree Project


UNIVERSITY OF SKÖVDE

# A CANONICAL CORRELATION ANALYSIS-BASED APPROACH TO IDENTIFY CAUSAL GENES IN ATHEROSCLEROSIS

Master Degree Project in bioinformatics
One year, 30 ECTS
Spring term 2018

Author:
Crisencia Sizyoongo (a17crisi@his.se)

External supervisors:
Ci Song (ci.song@igp.uu.se)
Marcel den Hoed (marcel.den_hoed@igp.uu.se)

Internal supervisor:
Björn Olsson (bjorn.olsson@his.se)
Examiner:
Zelmina Lubovac (zelmina.lubovac@his.se)

## Table of Contents

## Abbreviations and key words

| | |
|---|---|
| GWAS | Genome-wide association studies |
| Chr | Chromosome |
| CCA | Canonical Correlation Analysis |
| LDL | Low-density lipoprotein |
| HDL | High-density lipoprotein |
| Tg | Triglycerides |
| Tc | Total cholesterol |
| Glu | Glucose |
| MacLip_Area | Area where macrophages and lipid colocalise |
| NeuLip_Area | Area where neutrophils and lipid colocalise |
| YACCA | Yet Another Canonical Correlation Analysis |
| KCCA | Kernel Canonical Correlation Analysis |
| PEPD | PeptidaseD |
| TIMD*4* | T-cell immunoglobulin and mucin domain 4 |
| VEFA | Vascular endothelial growth factor A |
| VEFB | Vascular endothelial growth factor B |

## Abstract

Genome-wide association studies (GWASs) have identified hundreds of loci that are strongly associated with coronary artery disease and its risk factors. However, the causal variants and genes remain unknown for the vast majority of the identified loci. Zebrafish model systems coupled with clustered regularly interspaced short palindromic repeats-C–associated 9 (CRISPR Cas-9) mutagenesis have enabled the possibility to systematically characterize candidate genes in GWAS-identified loci. In this thesis, canonical correlation analysis (CCA) was used to identify putative causal genes in multiplexed genetic screens for atherogenic traits in zebrafish larvae in an efficient manner. The two datasets used in this thesis contained genes and phenotypes obtained through sequencing and high-throughput imaging of fish larvae. Dataset 1 contained (7 genes, 11 phenotypes, n = 384) and dataset 2 (4 genes, 11 phenotypes, n = 384). CCA's multiple genes vs. multiple phenotype analysis in dataset 1 identified the genes *met, pepd, timd4* and *vegfa* to have an association with the total cholesterol, triglycerides, glucose, corrected lipid disposition, as well as co-localization of (macrophage and lipid deposition,) (neutrophils and lipid deposition) and (macrophage and neutrophils). In dataset 2, CCA found previously reported correlation of genes *apobb1* and *apoea* with total cholesterol, low-density lipoprotein and triglycerides as well as co localization of neutrophils and lipids. In comparison with hierarchical linear model, CCA represents a powerful and promising tool to identify causal genes for cardiovascular diseases in data from zebrafish model systems.

## Introduction

Atherosclerosis plays a major role in the development of cardiovascular disease. It is a chronic inflammatory disease, elicited by the accumulation of lipids, inflammatory cells, smooth muscle cells, and extracellular matrix in the atrial intima [1]. The pathological aspects of atherosclerosis are well described, but the gene expression profiles during the atherosclerosis development are still largely unidentified [2]. The risk phenotypes associated with atherosclerosis include levels of low-density lipoprotein (LDL), high-density lipoprotein (HDL), triglycerides (Tg) and total cholesterol (Tc). These blood lipids levels are heritable, modifiable risk factors in coronary artery disease [3].

Genome-wide association studies (GWAS) are one of the major breakthroughs that have identified hundreds of loci in atherosclerosis. Findings from these studies have identified a locus on chromosome 9p21.3 (Chr9p21.3) to have the strongest genetic factor of atherosclerosis known to date [4-7]. The Global Lipids Genetics Consortium has identified and annotated 157 loci associated with the blood lipid levels, including 62 loci not previously associated with lipid levels in humans [3]. The major challenge post GWAS is to identify genetic variants and genes that are causal and integrate them into our understanding of the pathophysiology of this frequent disease. There is now a major interest in finding genetic variants or genes that are causal to the related phenotypes [8].

Recently, the zebrafish has become the trending animal model for investigating human genetic variants and diseases, supported by its genetic similarity to humans and outstanding manipulability [9]. The use of zebrafish and the CRISPR-cas9 technology has enabled to identify the causal genes in atherosclerosis [10]. Varsney et al (2015) developed a pipeline of generating multiple site directed mutations in zebrafish larvae and allow systematic screening of a large number of genes. A combination of up to eight single guide- RNA (one per gene) can be injected together into one-cell stage embryos to create loss-of-function mutations [11]. The genotype and phenotype information is thereby obtained using high-throughput sequencing and imaging on this mutated fish.

To identify causal genes, analytical methods and techniques play an important role. Methods such as Hierarchical Linear Models (HLM) could be used to find the independent effect of mutations in multiple targeted genes simultaneously. HLM is a complex form of ordinary least squares (OLS) regression that is used to analyze variance in the outcome variables when the predictor variables are at varying hierarchical levels [12]. Another method is multiple linear regression analysis that has been widely used in finding associations between a set of independent variables and dependent variables [13]. However these methods are prone to type-1 and type-2 errors resulting from multiple testing and lack the ability to pick up genes that have a modest effect on multiple outcomes [14-15].

In this thesis, a canonical correlation analysis (CCA) method is applied to identify the causal genes in atherosclerosis using the two variable sets consisting of genotype

and phenotype data. CCA is a method for examining the multivariate relation(s) between two sets of variables, with each consisting of two or more variables [16]. CCA has been proposed to provide an efficient and powerful approach for both univariate and multivariate tests of association [17]. CCA can be used to capture the underlying genetic background of a complex disease by associating two datasets containing information about a patient's phenotypic and genetic details, and has the ability to detect weak associations [18].

CCA is a commonly used method for data integration and it has been used in several studies [19-20]. The application of CCA for association analyses was initially proposed by Ferreira and Parcel [21]. A gene-based test of association using CCA done by Tang et al (2012), showed that including all genetic variation assayed in a given gene can be counterproductive when the specific aim of the analysis is to identify genes that harbor either uncommon or common causal variants, but not both [22]. Based on these results, they suggested that CCA might provide a particularly useful gene-based approach for the separate analysis of either uncommon or common variants. In another study, Na Yu et al (2017) applied CCA to study the relationships between anthropometric parameters and physical activity with blood lipids.

In their study, blood lipids showed a significant but moderate association with physical and anthropometric parameters. Waist circumference, BMI and occupational physical activity function were identified as major influences on lipids. They concluded that CCA is an efficient method to find the most influential factors on exposures and outcomes [23]. Further applications of CCA studies include: 1) analysis of blood phenotypes related with metabolic syndrome in mice [24]; and 2) use of CCA for single gene vs. multiple traits analysis to find different child behavior profiles [25].

## Problem definition

Analytical methods that involve multiple testing such as HLM are known to be prone to type -1 error. Therefore CCA method can be employed to overcome such problems. There is a lack of research applying CCA to study atherosclerosis development and the correlations between multiple genotypes and phenotypes for this disease. Existing results from other domains seem promising in providing more insight compared to HLM, and it is of interest to add to current research by comparing the results from HLM and CCA on genotype and phenotype data from a well-controlled animal model system of atherosclerosis.

## Aim and Objectives

The aim was to apply the CCA method approach using genotype and phenotype data in atherosclerosis. The objectives were to perform;

- CCA on datasets containing multiple genotypes and multiple phenotypes
- CCA on datasets containing single gene versus multiple phenotypes
- Compare the performance of CCA and HLM based on the results obtained using both methods

## Materials and Methods

### Datasets

The datasets used in this study have been generated using the zebrafish model system coupled with CRISPR-cas9 technology systems covering atherosclerosis. CRISPR technology has made it easier to both engineer specific DNA edits and to perform functional screens to identify genes involved in a phenotype of interest [26].

In total, two datasets were used. Dataset 1 (7 genes, 11 phenotypes, n = 384) contains previously unanticipated genes, which are being screened for causality from the GWAS [3]. Dataset 2 contain 4 known genes that have been studied as a proof of principal to if the zebrafish can be used as a model system to study atherosclerosis. Here n represents the number of fishes. Each dataset contain both the genotype and phenotype data. The genotype information was obtained through the sequencing of the CRISPR-cas9 target site using paired-end sequencing on a MiSeq (2x250 bp) whereas the phenotype data was obtained through the imaging of the traits of interest, in combination with enzymatic assays to quantify whole-body lipid and glucose levels. The whole-body lipid profile includes low density lipoprotein (LDL), high-density lipoprotein (HDL), triglycerides (Tg) and total cholesterol (Tc). The vascular atherogenic traits include vascular accumulation of lipids, macrophages (Mac), neutrophils (Neu) and their two-way co-localization (MacNeu, NeuLip and MacLip). All image-based atherogenic traits have been quantified from optical sections using custom-written scripts in publicly available tools [27, 28].

### The study design

The study is carried out in three stages. In the first stage CCA is performed on multiple genes vs. multiple phenotypes (in which the algorithm selects the sets of both genes and phenotypes that correlate); the second stage involves applying CCA on single gene vs. multiple phenotypes; the third stage is comparison of CCA and HLM performance.

### Canonical Correlation Analysis

CCA is a multivariate statistical model that aid in the study of linear interrelationships between two sets of variables by measuring the correlation between them, to quantify the strength of the relationship [29]. The simplest general formula for CCA is shown in Equation 1 where $X_1 \ldots X_n$ are independent and $Y_1, \ldots, Y_n$ are the dependent variables [30].

$$X_1 + X_2 X_3 + \cdots X_n = Y_1 + Y_2 + Y_3 + \cdots + Y_n \quad (1)$$

A composite score is formed from the linear combinations of two or more variables from a dataset called the canonical variates $\boldsymbol{\omega} = \boldsymbol{Y_u}$ and $\boldsymbol{\varepsilon} = \boldsymbol{X_v}$, with the weight vectors $\mathbf{u'} = (u_1,\ldots,u_p)$ and $\mathbf{v'} = (v_1,\ldots, v_q)$. The optimal weight vectors are obtained by maximizing the correlation between the canonical variate pairs, also known as the canonical correlation [31]. CCA develops a canonical function that maximizes the canonical correlation coefficient between the two canonical variates. The canonical correlation coefficient measures the strength of the relationship between the two canonical variates. The canonical function then maximizes the correlation between the two composite variables [32]. Further CCA develops as many functions as there are variables in the smaller variable set. Each function is orthogonal from the others so that they represent different relationships among the sets of dependent and independent variables [33]. The variables found in the first iteration of the method give the first of canonical variables [34]. The loadings of the individual variables differ in each canonical function and represent variables' contributions to the specific relationship being investigated. Now the challenge is to choose how many of them should be interpreted. In most cases the first function is the most legitimate. Loadings are correlations between the original variables in each set and their respective canonical variates [35]. Hair et al (2006) suggested three criteria of choosing the important functions, as they believed that the use of a single criterion such as the level of significance is too superficial. The three criteria are: (i) level of significance (ii) magnitude of the canonical correlation, and (iii) redundancy measure for the percentage of variance accounted for from the two datasets, like R2 statistic for multiple regression. No generally accepted guidelines have been established regarding suitable thresholds for canonical correlations values [36].

CCA provides a convenient statistical framework to simultaneously test the association between any number of quantitative phenotypes ($p$, phenotype-set) and any number of single nucleotide polymorphisms (SNPs) ($q$, SNP-set) genotyped across a gene or region of interest in unrelated individuals. The test is equivalent to (*i*) univariate linear regression when $p=q=1$ (single trait versus single SNP) and standard multiple regression when (*ii*) $p=1$ and $q>1$ (single trait versus multiple SNPs) or (*iii*) $p>1$ and $q=1$ (multiple traits versus single gene, considered by [37]. When (*iv*) $p>1$ and $q>1$ (multiple traits versus multiple genes), this approach represents a multivariate gene-based test of association [38]. Consider the $n \times p$ matrix **Y**, containing $p$ (Phenotype) variables and the $n \times q$ matrix **X**, containing $q$ (Genotype) variables, obtained from $n$ fish. CCA tries to find linear combinations of all the variables in **Y**, which correlate maximally with linear combinations of all the variables in **X.**

## Interpretation of Canonical Correlation Analysis

The common practice is to first analyze functions whose canonical correlation coefficients are statically significant beyond pre-agree level, typically 0.05 or less. Here, the three criteria recommended by Hair et al (2006) were followed: (i) Level of statistical significance of the function, (ii) magnitude of the canonical correlation, and (iii) redundancy measure for the percentage of variance accounted for from the

two datasets. The CCA test of significance of all canonical correlations is based on the Wilk's Lambda and Rao's F approximation [34], calculated as shown in equations 2-6. The Wilk's Lambda, λ, is calculated as:

$$\lambda = \prod_{i=1}^{j}(1 - c_i^2) \quad (2)$$

Where *j* is the number of canonical components, *c*, that are being calculated.

Rao's F approximation is defined as:

$$F(df1, df2) = \left(\frac{1-\lambda^{\frac{1}{2}}}{\lambda^{\frac{1}{2}}}\right) \times \left(\frac{df2}{df1}\right) \quad (3)$$

$$s = \sqrt{\frac{p^2 \; x \; q^2 - 4}{p^2 \; x \; q^2 - 5}} \quad (4)$$

$$df1 = p \; x \; q, \quad (5)$$

$$df2 = \left(n - 1.5 - \frac{p + q}{2}\right) x \left(s - \frac{p \; x \; q}{2} + 1\right) \quad (6)$$

Where $q$ is the number of SNPs in the genotype, $p$ the number of phenotypes evaluated, and *n* the number of samples.

CCA can be used for both continuous and categorical data of either dependent or independent variables [30]. To determine the relative importance of each original variable to each function three methods have been proposed (i) canonical weights (standardized coefficients), (ii) canonical loadings (structural correlations) and (iii) canonical cross-loadings. As the canonical weights, like regression weights, are vulnerable to multi collinearity, most of the literature suggests using canonical loadings or crossing loadings [39]. In this thesis both loadings and cross-loadings have been used to interpret the canonical variates. However, there is no established cut off.

A rule of thumb is that a variable loading is $\geq 0.30$ highlights an important contributing variable in to the function [40]. Structural loading or loading matrix is a matrix of correlations between the canonical coefficients and the variables in each set. It is created by multiplying the matrix of correlations between variables with the matrix of canonical coefficients [41].

## Implementation

All data were analyzed using R *https://www.rstudio.com*. The packages 'cca' and 'yacca' (yet another canonical correlation analysis) were used to achieve the CCA analysis. Initially CCA is performed with a built-in function called 'cancor' from the cca package. The CCA performs a canonical correlation (and canonical redundancy) analysis on two sets of variables but has limited output. For example, it does not contain *F*- test measures and plots. On the other hand the 'yacca' package provides

an alternative canonical correlation/redundancy analysis function, with associated print, plot, and summary methods. The cca and yacca packages have been previous used together [32,42]. The cancor2 function was therefore written in order to use both packages simultaneously, simplifying the interpretation of the CCA output. The cancor2 function returns two lists. List 1, which has cca output and list 2 has yacca output which includes the *F*-test, helio plots and summary report functions. When the cancor2 function was tested on the data, the output contained complex numbers, and then the modification was made to the yacca package in order to change the complex numbers to ordinary or real numbers.

## Alternative Methods

A non-linear based approach such as artificial neutral networks (ANNs) would be another alternative method to use in multivariate datasets. In this case, it can be employed to predict genetic variants or genes relationship with the phenotypes. ANNs are methods that are used for pattern recognition that have been widely used in the biology to solve a variety of problems [43]. They were originally developed to simulate activities and interactions among neurons in the brain. However, ANNs are now simply used as mathematical devices to model statistical relations. Lucek et al, (1997) developed the ANNs to identify sets of marker loci each linked to a disease locus conferring disease susceptibility. Neural networks can be designed to implement discriminant functions, which aim to classify sets of input values (independent variables) according to their associated output values (dependent variables), so that a given set of input values will produce a set of output*s* close to the observed values [44*].* Hsia et al (2003) used ANNs in the prediction of survival in surgical unresectable using genetic polymorphisms and clinical parameters in their dataset. In this study ANN achieved promising classification results when clinical parameters and genetic factors were considered simultaneously in the prediction model. In another study ANN was used to find associations between SNPs and childhood allergic asthma (CAA), which is more strongly influenced by genetic factors than other types of allergic asthma. [45]. They concluded that ANN could be used to characterize development of complex diseases caused by multiple factors.

ANNs have their limitations and problems. One of the limitations is the way a trained network makes its decisions. Because the information encoded by the network is just a collection of numbers, it is quite difficult to work out the reasoning that goes into its decision-making process. Neural networks are sometimes referred to as black boxes because of this limitation [46]. Another problem when using neural network is how to configure the network topology and appropriately control parameters. The connections between neurons in the neural network are very complicated and the configuration parameters of the network require much manual trial and error in order to gain an appropriate and functional network [47]. CCA was opted for instead because it is a linear method used in finding correlation in multivariate datasets.

## Results

The results are divided into i) multiple genes vs. multiple phenotypes (in which the algorithm selects the sets of both genes and phenotypes that correlate); ii) single gene vs. multiple phenotypes (in which the algorithm identifies the set(s) single gene associated with multiple phenotypes); and iii) comparison of results based on CCA and HLM approaches.

### DATASET 1

### Multiple genes vs. multiple phenotypes

CCA develops as many functions as there are variables in the smaller variable set. Table 1 shows seven functions because the independent set contained the minimum number of seven variables (genes). The correlations for each successive function (CVs) were 0.469, 0.353, 0.341, 0.286, 0.235, 0.165, and 0.097. The first correlation was statistically significant (p = 0.004, F-test) and the redundancy index was above zero (0.12) Figure 1. Therefore CV 1 is the only function noteworthy in this context.

**Table 1**. F Test for Canonical Correlations (Rao's F Approximation) for multiple genes vs. multiple phenotypes. Table 1 shows the canonical variate (CV) 1 to 7, correlation, F-statistics, number of degrees (Num den), degrees of freedom (Den df) and significant level (p -value)

| Canonical Variate (CV) | Correlation | F-statistics | Num den | Den df | P- value |
|---|---|---|---|---|---|
| **CV 1** | 0.469 | 1.493 | 77.000 | 948.25 | 0.004947 ** |
| **CV 2** | 0.353 | 1.208 | 60.000 | 832.87 | 0.139555 |
| **CV 3** | 0.341 | 1.116 | 45.000 | 714.35 | 0.281377 |
| **CV 4** | 0.286 | 0.923 | 32.000 | 591.65 | 0.590034 |
| **CV 5** | 0.235 | 0.736 | 21.000 | 462.86 | 0.795520 |
| **CV 6** | 0.165 | 0.509 | 12.000 | 324.00 | 0.908060 |
| **CV 7** | 0.097 | 0.313 | 5.0000 | 163.00 | 0.904272 |

Significant codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

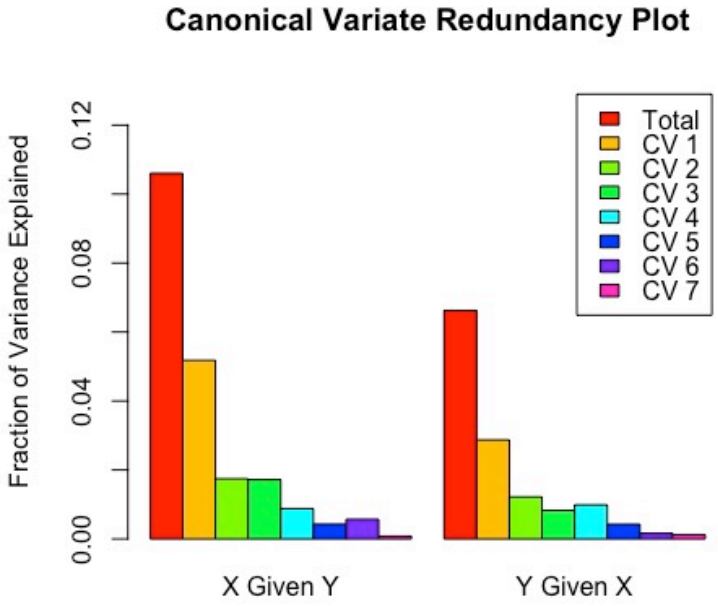


**Figure 1.** The redundancy plot of all the canonical variates (CVs). The X represents the X variables (genes) and Y represents the Y variables (phenotypes). The plot shows the fraction of variance explained in X given Y and in Y given X.

To further understand the contents in CV 1 function, the loadings and cross loadings of the variables for the 1st canonical function are presented in Table 2. Following the criterion recommended by Kabir et al (2014) the loading above 0.30 was considered important. Looking at the loadings of the variables for function (CV 1) the most influential variables are indicated by the asterisk (*). In the genotypes these include the genes *timd4 (loading:* -0.806), *met (*-0.676), *pepd (*-0.506*) and vegfabmm (*0.473). In the phenotypes, NeuLip_Area (-0.574), Cld (-0.513), Tg (-0.459), MacNeu_Area (0.472, Tc (-0.371), MacLip_Area (-0.328) and Glu (0.303) were the most influential factors. In multiple genes vs. multiple phenotypes, CCA analysis has indicated that the genes *timd4, met, pepd, and vegfb* have an association with the phenotypes NeuLip_Area, Tg, MacNeu_Area, MacLip_Area and Glu. The helio plot in Figure 3 **A** shows how the genotypes (X) are correlated with the phenotypes (Y). The relationship of these two set of variables is also visualized with the scatter plot **B** with moderate correlation of r = 0.4.

**Table 2.** Multiple genes vs. multiple phenotypes. The first column represents the variables (genotypes and phenotypes variables), the second column is the loadings of each variable and the third represent the crossing loadings of each variable.

| Variables | Loadings | Cross loadings |
|---|---|---|
| **Genotypes** | | |
| *map3k1* | -0.239 | -0.112 |
| *met* | -0.676* | 0.317 |
| *pccb* | -0.041 | -0.019 |
| *pepd* | -0.506* | -0.237 |
| *timd4* | -0.806* | -0.378 |
| *vegfaa* | 0.004 | 0.002 |
| *vegfab* | -0.473* | -0.222 |
| **Phenotypes** | | |
| LDL | 0.187 | 0.088 |
| HDL | -0.170 | -0.079 |
| Tc | -0.371* | -0.174 |
| Tg | -0.459* | -0.215 |
| Glu | 0.303* | 0.142 |
| Mac_Area | 0.075 | 0.035 |
| Neu_Area | 0.002 | 0.001 |
| Cld | -0.513* | -0.241 |
| MacLip_Area | -0.328* | -0.153 |
| NeuLip_Area | -0.574* | -0.269 |
| MacNeu_Area | 0.472* | 0.221 |

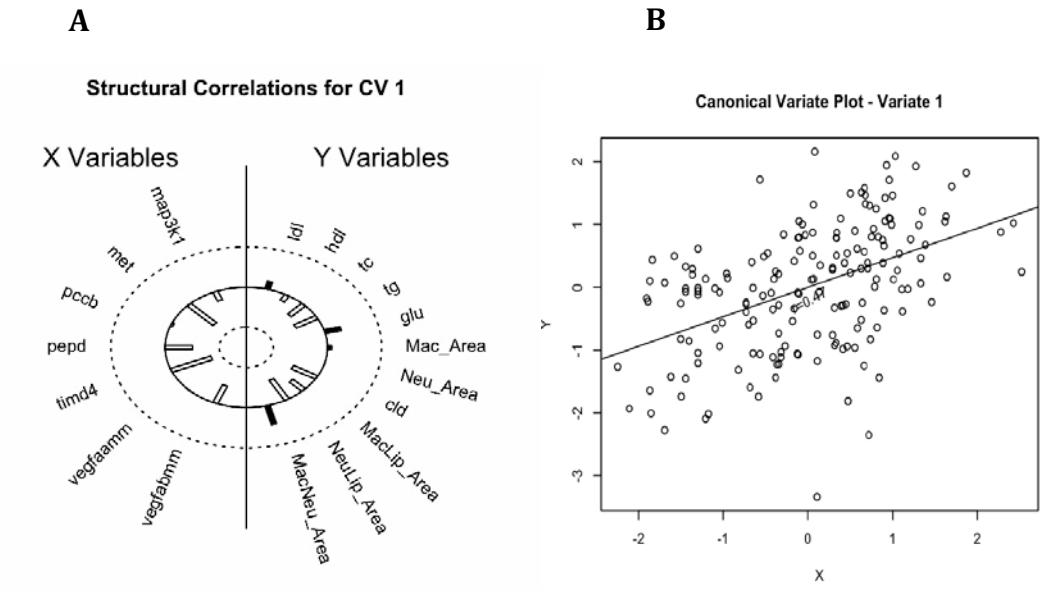


**Figure 2**. **A.** Helio plot of multiple genes vs. multiple phenotypes. The relative size of structural correlation is indicated by the relative length of the bars is from 0 (solid line) to 1 (dotted lines). The bar extending towards the circumference indicate (positive correlation) and towards the center (negative correlations). **B.** Scatter plot of Y variables and X variables with correlation r = 0.4

## Single gene vs. multiple phenotypes

To find the correlation between single gene vs. multiple phenotypes CCA was applied. Following the same rule applied in multiple genes vs. multiple phenotypes, the loading of 0.30 and above was considered as the influential variable Table 3.

**Table 3**. Single gene vs. multiple phenotypes. The first column represents the gene, the second column represents the correlation, the third column represents the p-value and the fourth column represents the phenotypes association. In the parenthesis is the single phenotype loading value.

| Gene | Correlation | P-value | Phenotype |
|---|---|---|---|
| ***map3k1*** | 0.293 | 0.178 | LDL (-0.445)<br>Tg (0.418)<br>Glu (0.312)<br>Cld (0.510)<br>MacLip_Area (0.595)<br>NeuLip_Area (0.5123) |
| ***met*** | 0.365 | 0.012 * | Tc (-0.395)<br>Tg (-0.610)<br>Mac_Area (0.359)<br>Cld (-0.547) |

| | | | |
|---|---|---|---|
| | | | MacLip_Area (0.428)<br>NeuLip_Area (-0.636) |
| ***pccb*** | 0.281 | 0.245 | Tg (-0.370)<br>Glu (-0.590)<br>Cld (-0.466)<br>MacLip_Area (0.656) |
| ***pepd*** | 0.314 | 0.095. | Cld (-1.2e-01)<br>NeuLip_Area (-4.3e-01) |
| ***timd4*** | 0.307 | 0.001 ** | LDL (0.490)<br>Tg (-0.374)<br>Cld (-0.391)<br>MacNeu_Area (0.523) |
| ***vegfa*** | 0.296 | 0.1662 | HDL (-0.325)<br>Mac_Area (0.373)<br>NeuLip_Area (-0.341)<br>MacNeu_Area (0.578) |
| ***vegfb*** | 0.307 | 0.1182 | LDL (0.490)<br>Tg (-0.374)<br>Cld (-0.391)<br>MacNeu_Area (0.523) |

The outcomes of the single gene of multiple phenotypes Table 3 include; the correlation p-value of CCA association, the phenotypes and their loadings. 7 genes have indicated an association with the multiple phenotypes. Gene *met* has shown to be correlated with Tc, Tg, Mac_Area, Cld, MacLip_Area and NeuLip_Area (p-value = $1.4 \times 10^{-2}$), *timd4* is correlated with LDL, Tg, Cld and MacNeu_Area (p-value = $1.7 \times 10^{-3}$). Although not significantly correlated, the gene *pepd* has shown association with Cld and NeuLip_Area (p-value = 0.09). The genes *map3k1, vegfa* and vegfb also indicated association with LDL, Tg, Glu, Cld, MacLip_Area and NeuLip_Area (p-value = 0.1), HDL Mac_Area, NeuLip_Area and MacNeu_Area and (DL, Tg, Cld and MacNeu_Area respectively. Despite *map3k* showing an overall weak association with the phenotypes (p-value = 0.2). Individual phenotype's, MacLip_Area, Cld, NeuLip_Area, Tg revealed moderate positive correlation (0.60, 0.51, 0.051 and 0.42) respectively Figure 4. They also cluster together which might indicate that mutation in *map3k* affects the phenotypes in a similar fashion. The phenotypes MacNeu_Area and LDL have also shown moderate correlation with the gene *vegfb* and timd4.

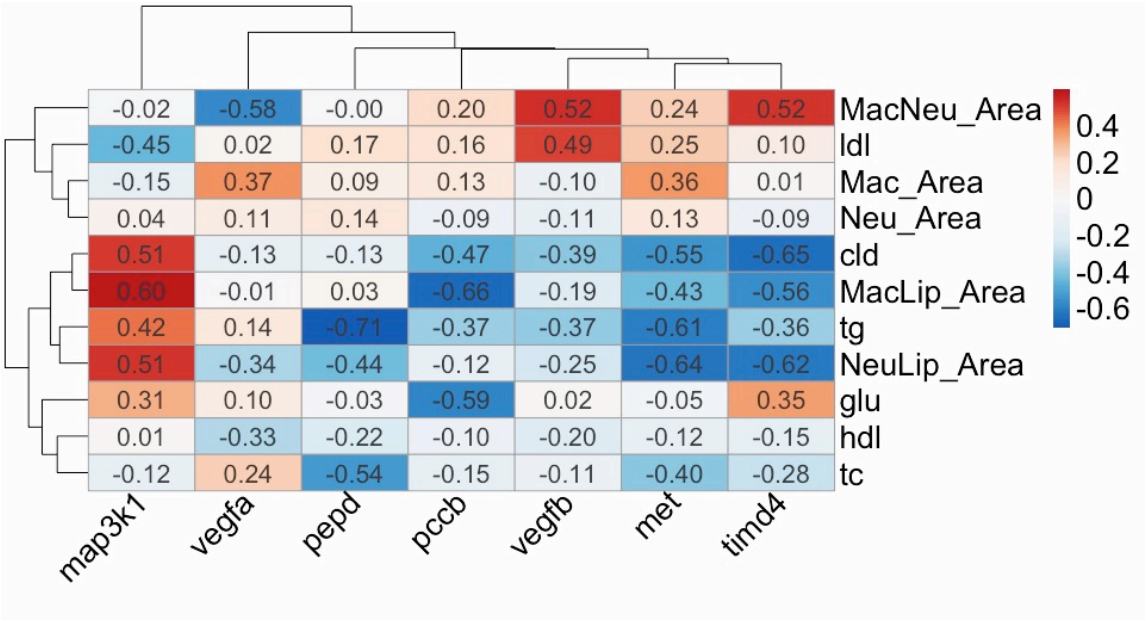


**Figure 3.** Heatmap single genes vs. multiple phenotypes cross correlation. The row is the phenotypes and the column is the gene from scale of (-0.6 to 0.4). The heatmap is generated using the values obtained in the cross correlation (CCA output). The values in the heatmap represent the direct measure of the gene-phenotype correlation relationships. The clustering is based on Euclidean clustering method.

In addition negative correlation was also observed with some phenotypes for example *pepd* was strongly negative correlated (-0.71) with the Tg and moderate correlation with Tc (-0.54). pccd with NeuLip_Area (-0.64) and Tg (-0.61). Lastly timd4 showed negative correlation with Cld (-0.65), and NeuLip_Area (-0.62).

The explained variance of single gene vs. multiple phenotypes is shown in figure 5. By definition explained variation measures the proportion to which a mathematical model accounts for the variation (dispersion) of a given dataset. Often, variation is quantified as variance [48]. That is the $R^2$ is the proposition of the variation in y that can be explained by the linear relationship between x and y. In Figure 5 shows 50% of the variation in the gene *pepd* can be explained by the linear relationship between Tg and *pepd* and *40% variation is explained between pccd* and Maclip_area and so on. The phenotype with high-explained variance forms a cluster (Cld, MacLip, Tg and NeuLip_area). These phenotypes have been identified in the multiple genes vs. multiple phenotypes as some of the most influential variables driving the correlation.

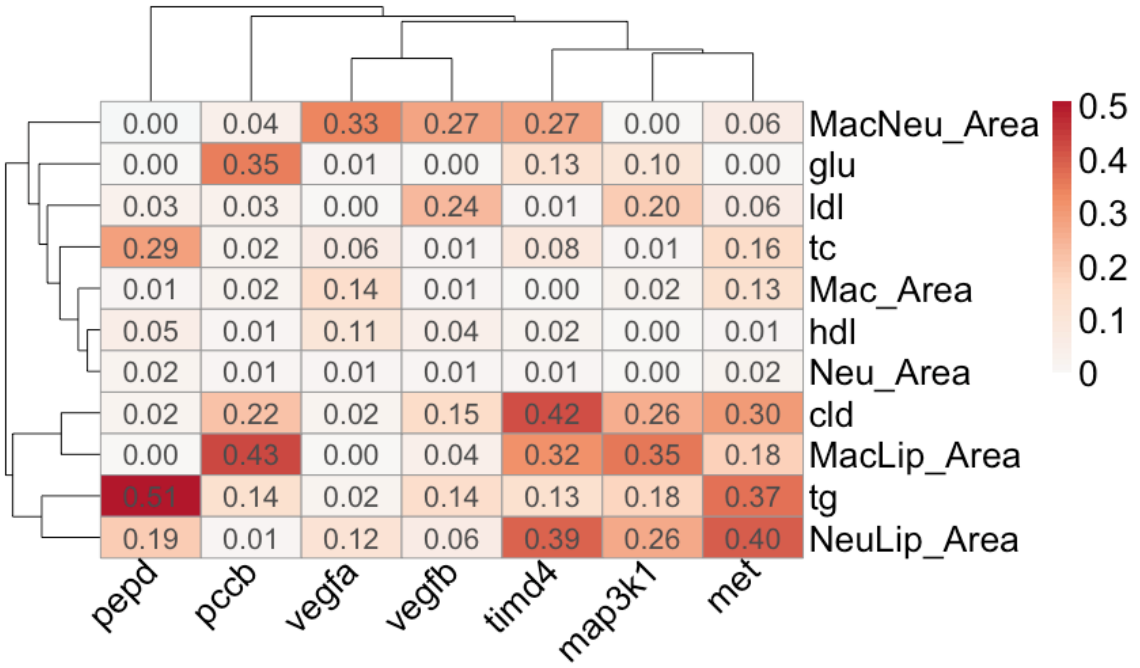


**Figure 4**. Heatmap of single gene vs. multiple phenotype explained variance. The row represents the phenotypes and the genes are represented in the column from the scale of (0 to 0.50).

## DATASET 2

### Multiple genes vs. multiple phenotypes

CCA was applied in the same way as in dataset 1. No Canonical variate was found significant; the selection of the most important CV was based on the correlation magnitude size in this case CV 1 with correlation of 0.33.

**Table 4.** F Test for Canonical Correlations (Rao's F Approximation) for multiple genes vs. multiple phenotypes. Column 1 represents the canonical variate (CV) 1 to 4, Column 2 represents correlation, Column 13represents F-statistics, Column 4 represents number of degrees (Num den), Column 5 represents degrees of freedom (Den df) and Column 6 represents significant level (p -value)

| Canonical Variate (CV) | Correlation | F | Num df | Den df | Pr (>F) |
|---|---|---|---|---|---|
| **CV 1** | 0.334 | 1.239 | 44.000 | 839.9 | 0.140 |
| **CV 2** | 0.292 | 0.914 | 30.000 | 646.4 | 0.599 |
| **CV 3** | 0.256 | 0.405 | 18.000 | 442.0 | 0.986 |
| **CV 4** | 0.089 | 0.222 | 8.000 | 222.0 | 0.986 |

To find out the major contributing variables in each set, the loadings and cross loadings are presented in table 5. The most influential variables are indicated by asterisk *. The genes *apobb1, apoea, apoeb* with loadings of -0.709, 0.706 and -0.220 contribute the most in the independent variables respectively. Tc, Mac_Area, Neu_Area, Cld and MacNeu_Area with loadings 0.386, 0.316, -0.316,0.302, 0.307 are the most influential factors or variables in the dependent set.

**Table 5.** Multiple genes vs. multiple phenotypes in CV 1 .The first column represent the variables (independent and dependent), the second column is the loadings of each variable and the third represent the crossing loadings of each variable.

| Variables | Loadings | Cross loadings |
|---|---|---|
| **Genotype** | | |
| *apobb1* | -0.709* | -0.013 |
| *apoea* | 0.706* | -0.009 |
| *apoeb* | -0.220* | -0.083 |
| *LDLra* | -0.000 | 0.0352 |
| **Phenotypes** | | |
| LDL | 0.252 | 0.084 |
| HDL | 0.0297 | 0.009 |
| Tc | 0.386* | 0.129 |
| Tg | 0.030 | -0.010 |
| Glu | 0.055 | 0.018 |
| Mac_Area | 0.316* | 0.105 |
| Neu_Area | -0.316 | -0.105 |
| Cld | 0.302* | 0.101 |
| MacLip_Area | 0.228 | 0.076 |
| NeuLip_Area | -0.068 | -0.022 |
| MacNeu_Area | 0.307* | 0.103 |

## Single gene vs. multiple phenotypes

**Table 6**. Single gene vs. multiple phenotypes association. The first column represents the gene, the second column represents the correlation, the third column represents the p- value and the fourth column represents the phenotypes association. In the parenthesis is the single phenotype loading value

| Gene | Correlation | P-value | Phenotype (loadings) |
|---|---|---|---|
| ***apobb1*** | 0.287 | 0.05183. | Tc *(0.562)*<br>Mac_Area *( 0.457)* |
| ***apoea*** | 0.278 | 0.07606 . | LDL *(0.345)*<br>Tg *( 0.318)*<br>Cld *(0.335)* |
| ***apoeb*** | 0.12467 | 0.9813 | LDL *( -0.402)*<br>Mac_Area *(0.505*)<br>Cld (*0.315)* |
| ***LDLra*** | 0.2595 | 0.1492 | Mac_Area ( -0.550)<br>Tc *(-0.423)*<br>HDL *(-0.389)* |

The gene *apobb1* has *shown to be correlated with the Tc and Mac_Area (p = 0.05).* The gene *apoea* also indicated positive correlation with *LDL, Tg, and Cld.* Apoeb and apoea are paralogs and both genes show correlation with LDL and Cld. Numerous studies have established that there is a striking correlation between *Apoe* and the LDL. The study by Nakashima et al. (1994) showed that APOE-deficient mice develop severe atherosclerosis due to increased circulating LDL [49]. The gene LDLra also has shown correlation with Mac-Area, Tc and HDL. Although not significant the correlation of LDLra with Mac_Area, Tc and HDL is in line with previous studies that have shown deficiency in the receptor LDL (LDLra) is major cause of familial hypercholesterolemia in humans [50].

## Comparison of results based on CCA and HLM approaches

The current standard method used to find associations of genotype and phenotype is HLM that is based on ordinary least squares. The performance of CCA and HLM was evaluated by assessing the associations found by both methods Table 7.

**Table 7.** Comparison of associations found by CCA and HLM. The first column represents the genes; the second column represents the phenotypes found by CCA with the corresponding gene. The third column represent the phenotypes found by HLM with the associated gene. Dark red represents significant associations.

| | CCA Associations | HLM Associations |
|---|---|---|
| Genes | **Phenotypes** | **Phenotypes** |
| ***map3k1*** | MacLIp_Area<br>Cld<br>NeuLip_Area | |
| ***met*** | **Tg**<br>**MacLip_Area**<br>**Mac_Area**<br>**Tc**<br>**Cld**<br>**NeuLip_Area** | **Tg**<br>**MacLip-Area**<br>**Mac_Neu** |
| ***pccb*** | MacLIp-Area<br>Glu<br>Cld | **MacLip_Area**<br>**NeuLIp_Area** |
| ***pepd*** | **Tg**<br>**Tc**<br>**NeuLip_Area** | |
| ***timd4*** | **MacLip_Area**<br>**Cld**<br>**NeuLip_Area**<br>**MacNeu_Area** | ***MacLip_Area***<br>***Glu*** |
| ***vegfa*** | NeuLip_Area<br>MacNeu_Area | **NeuLip_Area** |
| ***vegfb*** | LDL<br>MacNeu_Area | **LDL** |

Both methods found identical phenotypes to be significantly correlated with the genes in most cases. Tg and MacLip_Area were found to be correlated with the *met* and MacLip_Area with *timd4* in both methods. Although not significant in CCA but significant in HLM, NeuLip_Area and LDL were shown to be correlated with *vegfa* and *vegfb* respectively. Overall CCA found more phenotypes than HLM in the case were both genes were significantly correlated with the multiple phenotypes. This can be seen in the gene met and *timd4* were additional phenotypes were found compared to HLM.

**Table 8.** Comparison on the CCA and HLM found (dataset 2). The first column represents the genes; the second column represents the phenotypes found by CCA with the corresponding gene. The third column I represent the phenotypes found by HLM with the associated gene. Dark red represents significant associations.

| | **CCA associations** | **HLM associations** |
|---|---|---|
| Gene | **Phenotypes** | **Phenotypes** |
| ***apobb1*** | Tc<br>Mac_Area | Cld<br>NeuLip_Area |
| ***apoea*** | LDL<br>Cld<br>Tg<br>MacLip_Area<br>NeuLip_Area | LDL<br>Cld |
| ***apoeb*** | LDL<br>Mac_Area<br>Cld | MacLip_Area<br>NeuLIp_Area<br>MacNeu |
| ***LDLra*** | Tc<br>Mac_Area<br>HDL | |

The comparison CCA and HLM results Table 8 show that both methods have identified the genes *apobb1* and *apoeb* to be significantly correlated with the phenotypes associated. In both cases similar phenotypes have been found, for example LDL and Cld are correlated with apoea. In additional CCA has found *apoea* to be correlated with Tg, co-localization of macrophages and lipid and neutrophils and lipids.

## Discussion

CCA was applied to study the association of genes associated with phenotypes in the development of atherosclerosis. CCA uses information from all the variables in the exposure and outcome variable sets and maximizes the estimation of the relationship between the two sets [23]. In multiple genes vs. multiple phenotypes analysis, CCA shed light on the overall correlation between a set of genes with a set of phenotype in dataset 1. CCA identified the genes *met, pepd, timd4* and *vegfb* to have an association with the phenotypes Tc, Tg, Glu, Cld, MacLip_Area, NeuLip_Area and MacNeu_Area.

In single gene vs. multiple analyses, CCA showed that *timd4* is correlated with LDL, Tg, Cld and MacNeu. Studies have shown that a single nuclear polymorphism in the TIM-4-encoding gene timd4 is associated with lowered LDL, triglycerides, and

cardiovascular disease [51 - 52]. Another study showed that *tim-4* mRNA negatively correlated with LDL levels in mice having type 2 diabetes mellitus. Folks et al (2016) also showed that blockade of *timd4* enhance atherosclerosis in LDL [53]. The gene *vegfa* and *vegfb* have shown to be correlated with LDL and the co-localization of neutrophils and lipids. Previous studies have shown that vascular endothelial growth factors (vegfs) participate in atherosclerosis, arteriogenesis, cerebral edema, neuro-protection, neurogenesis, angio- genesis, post ischemic brain and vessel repair, and the effects of transplanted stem cells in experimental stroke [54]. The role of *vegfs* in atherosclerosis has been tested in a variety of animal models. In LDL receptor-knockout mice fed an atherogenic diet, developed hyperlipidemia and atherosclerosis. Vaccination against vegf receptor 2 (vegfr-2/flk-1) reduced the size and micro-vessel density of aortic atherosclerotic lesions [55]. However more functional studies are needed to understand the regulation and expression of these genes on transcription level.

In dataset 2 the proof of principal study, the single gene vs. multiple phenotype analysis showed interesting result Table 6 which are consistent with previous studies. For example *apoea* is associated with LDL [56], *apobb1* with HDL and Tc. LDL, HDL and total cholesterol. These lipids play an important role in the lipid metabolism and development of atherosclerosis [57]. The results found indicate that zebrafish can be used as a model for cardiovascular and metabolic diseases such as atherosclerosis.

The performance of CCA and HLM was assessed through the correlations of single gene vs. multiple phenotypes found. Both methods identified *met and timd4* to be significantly correlated with Tg and MacLip_Area and MacLip_Area respectively. Although significant in HLM and not in CCA, *vegfa* and *vegfb* were shown to have association with NeuLip_Area and LDL. It has been observed that CCA tends to found more phenotypes than HLM in the case were both genes were significantly correlated with the multiple phenotypes. This can be seen in the gene *met* and *timd4* were additional phenotypes were found in CCA compared to HLM. CCA shows accumulative effects of individual associations. The ability to capture more phenotypes is also observed in dataset 2 were more phenotypes associated with the *apoea* were identified.

CCA analysis of multiple genes vs. multiple phenotypes could be useful for information gain on the overall relationship between the phenotypes and genotypes. In the case of single gene vs. multiple phenotypes analysis, the loadings in the phenotypes could be used as a feature selection step in univariate analysis such as HLM. Hence CCA could be used as a complimentary method to HLM in finding association in multivariate dataset. Because of limiting the inefficiencies that may accompany conventional multiple testing, CCA could help to reduce type-1 error (an error for refusing the truth, usually represented by "α") and add accuracy to its results [4]. A previous study of CCA has been shown to increase the statistical power

in detection of previously reported genetic associations, and identified a number of novel pleiotropic associations between genetic variants and phenotypes [58].

## Conclusion

In conclusion, CCA approach was applied to find the correlation between the genotypes and the phenotypes in atherosclerosis. The genes *met, timd4, pepd* and *pccd* have been identified to be the causal genes in the development of atherosclerosis. These genes have showed to be correlated with lipids Tc, Tg and the cells macrophage, neutrophils and co localization of macrophages and neutrophils. In terms of single gene vs. multiple phenotypes, *met* was significantly correlated with Tc, Tg, Mac_Area, Cld, MacLip_Area and NeuLip-Area and *timd4* was significantly correlated with LDL, Tg, Cld and MacNeu_Area. The identification of these genes will enhance our understanding of lipid metabolism in atherosclerosis pathophysiology and will aid in identifying drug target and treatment. The use of CCA method has indicated to be an effective method finding out the influential factors in both sets of variables and assess the association between the genotypes and the phenotypes in atherosclerosis studies. It may offer an efficient, practical and more biologically comprehensive approach to assessing the association between two sets of variables, by taking into account the innate complexity of interactions and biological pathways between variables. Consequently CCA can be used as first baseline of analysis in a multivariate datasets and a complimentarily method to HLM in optimizing and validation in the identification of causal genes in atherosclerosis studies.

## Limitations of CCA and Future Perspective

Among the limitations that can have the greatest impact on the results and their interpretation are the following: i) the canonical correlation reflects the variance shared by the linear composites of the sets of variables, not the variance extracted from the variables; ii) Canonical weights derived in computing canonical functions are subject to a great deal of instability; iii) Canonical weights are derived to maximize the correlation between linear composites, not the variance extracted; iv) The interpretation of the canonical variates may be difficult, because they are calculated to maximize the relationship, and aids for interpretation, such as rotation of variates in factor analysis, are limited. It is difficult to identify meaningful relationships between the subsets of independent and dependent variables because precise statistics have not yet been developed to interpret canonical analysis, and we must rely on inadequate measures such as loadings or cross-loadings [38]. Future research directions include improving the method, by developing an algorithm that can find the precise statistical value of the individual linear combinations. As it stands now only the correlation values are obtained in the pairwise combinations (single gene vs. single phenotype) results shown in figure 4, and it is difficult to interpret how significant are these correlations found. Trying out the non-linear

canonical correlation relations using kernel based approaches such as KCCA can also be interesting to analyse if that can improve the results. Then we can determine and conclude, if the type of data generated in Zebrafish model confers more to linear methods or non-linear methods.

## Impact of research on the society.

Cardiovascular disease is still the leading cause of death world- wide and is mainly caused by atherosclerosis and its thrombotic complications (59). With the increasing availability of genomic data, the use of statistical models that can provide efficient analysis in understanding of causal genes in atherosclerosis are needed. Therefore data analysis was performed using CCA over the datasets obtained in atherosclerosis. The impact of this study on the society is that the proposed method CCA has been shown to provide more insight on the risk factors involved in the development of atherosclerosis. The idea of comparing different methods over the same set of data also solidifies the findings and we can be certain moving forward in further research of drug treatment and prevention. In the screening of genetic variants associated with atherosclerosis, the genes met, *pccd, timd4* and *vegfb* have been identified as the causal genes of atherosclerosis. The genes identified and their associated phenotypes can be used as targets and predictive biomarkers in the therapeutic research. In summary this project has highlighted on how the CCA model works and the biological mechanisms in atherosclerosis development and has enhance our understanding of atherosclerosis.

## Ethics Statement

The data used in this paper was collected in line with the Swedish regulations, and all experiments have been approved by Uppsala Djurförsöksetiska nämnd, Uppsala, Sweden (Permit numbers C142/13 and C14/16.

## Acknowledgements

This research was carried out at Uppsala Sci-lifelab. (Uppsala University, 2018)

Special thanks to Assistant Prof. Marcel den Hoed for accommodating me in his research group and enable me carry out this research successfully. My sincere gratitude goes to my supervisors Björn Olsson (Skövde University), Marcel den Hoed and Ci Song (Uppsala University) for their tremendous supervision and guidance.

To my partner Roland thank you for being such a strong support system in understanding of my shortcomings and for the help when really needed. To the whole Marcel's group member's thank you for welcoming me and making my stay a worthy while.

## References.


1. Galkina, Elena, and Klaus Ley. "Immune and inflammatory mechanisms atherosclerosis." *Annual review of immunology* 27 (2009): 165-197
2. LuTgens, Esther, et al. "Gene profiling in atherosclerosis reveals a key role for small inducible cytokines: validation using a novel monocyte chemoattractant protein monoclonal antibody." *Circulation* 111.25 (2005): 3443-3452
3. Willer, Cristen J., et al. "Discovery and refinement of loci associated with lipid levels." *Nature genetics* 45.11 (2013): 1274.
4. Helgadottir, Anna, et al. "A common variant on chromosome 9p21 affects the risk of myocardial infarction." *Science* 316.5830 (2007): 1491-1493
5. McPherson, Ruth, et al. "A common allele on chromosome 9 associated with coronary heart disease." Science 316.5830 (2007): 1488-1491
6. Samani, Nilesh J., et al. "Genomewide association analysis of coronary artery disease." *New England Journal of Medicine* 357.5 (2007): 443-453
7. Wellcome Trust Case Control Consortium. "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls." *Nature* 447.7145 (2007):
8. Holdt, Lesca M., and Daniel Teupser. "From genotype to phenotype in human atherosclerosis-recent findings." Current opinion in lipidology 24.5 (2013):
9. Liu, Jiaqi, et al. "CRISPR/Cas9 in zebrafish: an efficient combination for human genetic diseases modeling." *Human genetics* 136.1 (2017): 1-12. .
10. Chadwick, Alexandra C., and Kiran Musunuru. "CRISPR-Cas9 Genome Editing for Treatment of Atherogenic Dyslipidemia Highlights." *Arteriosclerosis, thrombosis, and vascular biology* 38.1 (2018): 12-18.
11. Varshney, Gaurav K., et al. "High-throughput gene targeting and phenotyping in zebrafish using CRISPR/Cas9." *Genome research* 25.7 (2015): 1030-1042.].
12. Hofmann, David A. "An overview of the logic and rationale of hierarchical linear models." *Journal of management* 23.6 (1997): 723-744
13. O'Reilly, Paul F., et al. "MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS." *PloS one* 7.5 (2012): e34861
14. Zhang, Zhen, Michael J. Zyphur, and Kristopher J. Preacher. "Testing multilevel mediation using hierarchical linear models: Problems and solutions." *Organizational Research Methods* 12.4 (2009): 695-719
15. Moiseev, Nikita A. "p-Value adjustment to control type I errors in linear regression models." *Journal of Statistical Computation and Simulation* 87.9 (2017): 1701-1711
16. Akaike, Hirotugu. "Canonical correlation analysis of time series and the use of an information criterion." *Mathematics in Science and Engineering*. Vol. 126. Elsevier, 1976. 27-96
17. Tang, Clara S., and Manuel AR Ferreira. "A gene-based test of association using canonical correlation analysis." Bioinformatics 28.6 (2012): 845-850.
18. Bush, William S., and Jason H. Moore. "Genome-wide association studies." PLoS computational biology 8.12 (2012): e1002822
19. Hotelling, Harold. "Relations between two sets of variates." *Biometrika* 28.3/4 (1936): 321-377

20. Lê Cao, Kim-Anh, Ignacio González, and Sébastien Déjean. "integrOmics: an R package to unravel relationships between two omics datasets." *Bioinformatics* 25.21 (2009): 2855-2856.
21. Ferreira, Manuel AR, and Shaun M. Purcell. "A multivariate test of association." *Bioinformatics* 25.1 (2008): 132-133.
22. Tang, Clara S., and Manuel AR Ferreira. "A gene-based test of association using canonical correlation analysis." Bioinformatics 28.6 (2012): 845-850.]
23. Yu, Na, et al. "Canonical correlation analysis (CCA) of anthropometric parameters and physical activities with blood lipids." *Lipids in health and disease* 16.1 (2017): 236.
24. Maria, et al. "Discovering genetic associations with high-dimensional neuroimaging phenotypes: a sparse reduced-rank regression approach." *Neuroimage* 53.3 (2010): 1147-1159.
25. Mick E, McGough J, Loo S, Doyle AE, Wozniak J, et al. (2011) Genome-Wide Association Study of the Child Behavior Checklist Dysregulation Profile. *Journal of the American Academy of Child and Adolescent Psychiatry* 50: 807–817.
26. Smith, Caitlin. "Editing the editor: Genome editing gets a makeover with CRISPR 2.0." Science 355.6321 (2017): 210-210.].
27. Schindelin, Johannes, et al. "Fiji: an open-source platform for biological-image analysis." Nature methods 9.7 (2012): 67.6.
28. Carpenter, Anne E., et al. "CellProfiler: image analysis software for identifying and quantifying cell phenotypes." *Genome biology* 7.10 (2006): R100.
29. Johnson, Richard A., and Dean Wichern. Multivariate analysis. John Wiley & Sons, Ltd, 2002.
30. Hair, Joseph F., et al. "Multivariate Data Analysis: Pearson Prentice Hall." Upper Saddle River, NJ (2006).
31. Waaijenborg, Sandra, and Aeilko H. Zwinderman. "Correlating multiple SNPs and multiple disease phenotypes: penalized nonlinear canonical correlation analysis." *Bioinformatics* 25.21 (2009): 2764-2771.
32. Anderson, Marti J., and Trevor J. Willis. "Canonical analysis of principal coordinates: a useful method of constrained ordination for ecology." Ecology 84.2 (2003): 511-525.
33. Kabir, Alamgir, et al. "Canonical correlation analysis of infant's size at birth and maternal factors: a study in rural Northwest Bangladesh." *PloS one* 9.4 (2014): e94243.
34. Kettenring, Jon R. "Canonical analysis of several sets of variables." Biometrika 58.3 (1971): 433-451.
35. Mardia, K. V.; Kent, J. T.; and Bibby, J. M. 1979. Multivariate Analysis. London: Academic Press.
36. Bartlett, M. S. "The statistical significance of canonical correlations." *Biometrika* 32.1 (1941): 29-37.
37. Holmans, Peter, et al. "Gene ontology analysis of GWA study data sets

provides insights into the biology of bipolar disorder." The American Journal of Human Genetics 85.1 (2009): 13-24.
38. Tang, Clara S., and Manuel AR Ferreira. "A gene-based test of association using canonical correlation analysis." Bioinformatics 28.6 (2012): 845-850.
39. Liu, Jing, et al. "Examination of the relationships between environmental exposures to volatile organic compounds and biochemical liver tests: application of canonical correlation analysis." *Environmental research* 109.2 (2009): 193-199.
40. Lambert, Zarrel V., and Richard M. Durand. "Some precautions in using canonical analysis." Journal of Marketing Research 12.4 (1975): 468-475.
41. Johnson, Richard A., and Dean Wichern. *Multivariate analysis*. John Wiley & Sons, Ltd, 2002.
42. Waller, Tomasz, et al. "DNA microarray integromics analysis platform." BioData mining 8.1 (2015): 18.
43. Bhat, Ajita, Paul R. Lucek, and Jurg Ott. "Analysis of complex traits using neural networks." Genetic epidemiology 17.S1 (1999).
*44.* Bishop, Chris, and Christopher M. Bishop. *Neural networks for pattern recognition*. Oxford university press, 1995.
45. Tomita, Yasuyuki, et al. "Artificial neural network approach for selection of susceptible single nucleotide polymorphisms and construction of prediction model on childhood allergic asthma." *BMC bioinformatics* 5.1 (2004): 120
46. Hsia, Te-Chun, et al. "Prediction of survival in surgical unresectable lung cancer by artificial neural networks including genetic polymorphisms and clinical parameters." *Journal of clinical laboratory analysis* 17.6 (2003): 229-234.
47. Leverington, David. "A basic introduction to feedforward backpropagation neural networks." *Neural Network Basics*( 2009).
*48.* https://en.wikipedia.org/wiki/Explained_variation
49. Nakashima, Yutaka, et al. "ApoE-deficient mice develop lesions of all phases of atherosclerosis throughout the arterial tree." Arteriosclerosis, thrombosis, and vascular biology 14.1 (1994): 133-140.
50. Liu, Chao, et al. "Modeling hypercholesterolemia and vascular lipid accumulation in LDL receptor mutant zebrafish." *Journal of lipid research* 59.2 (2018): 391-399.
51. Kathiresan, Sekar, et al. "Common variants at 30 loci contribute to polygenic dyslipidemia." Nature genetics 41.1 (2009): 56.
52. Do, Ron, et al. "Common variants associated with plasma triglycerides and risk for coronary artery disease." Nature genetics 45.11 (2013): 1345.].
53. Foks, Amanda C., et al. "Blockade of Tim-1 and Tim-4 enhances atherosclerosis in low-density lipoprotein receptor–deficient mice." Arteriosclerosis, thrombosis, and vascular biology (2016): ATVBAHA-115.
54. Greenberg, David A., and Kunlin Jin. "Vascular endothelial growth factors

(VEGFs) and stroke." *Cellular and molecular life sciences* 70.10 (2013): 1753-1761.

55. Petrovan, Ramona J., et al. "DNA Vaccination Against VEGF Receptor 2 Reduces Atherosclerosis in LDL Receptor–Deficient Mice." *Arteriosclerosis, thrombosis, and vascular biology* 27.5 (2007): 1095-1100.
56. Liu, Chao, et al. "Modeling hypercholesterolemia and vascular lipid accumulation in LDL receptor mutant zebrafish." *Journal of lipid research* 59.2 (2018): 391-399

57. Davignon, Jean, Richard E. Gregg, and Charles F. Sing. "Apolipoprotein E polymorphism and atherosclerosis." Arteriosclerosis, Thrombosis, and Vascular Biology 8.1 (1988)

58. Seoane, Jose A., et al. "Canonical correlation analysis for gene-based pleiotropy discovery." PLoS computational biology10.10 (2014): e1003876.

59. Smith, Sidney C., et al. "Principles for national and regional guidelines on cardiovascular disease prevention: a scientific statement from the World Heart and Stroke Forum." *Circulation* 109.25 (2004): 3112-3121.