

# University of Ottawa Department of Biology

#### HONOURS B.Sc. BIOMEDICAL SCIENCE, OPTION IN BIOSTATISTICS

## Development and Testing of an Optimal Cardiometabolic Genetic Risk Score to Predict Coronary Artery Disease Risk

Honours Dissertation of:

Christopher B. Cole

Thesis Supervisor:

Prof. Ruth McPherson, MD, PhD, FACP, FRCPC, FRSC

Secondary Thesis Supervisor:

Dr. Majid Nikpay, PhD

### Preface

Fill in later

CHRISTOPHER B. COLE
Ottawa
May 2016

#### Abstract

Background and Rationale: Coronary artery disease (CAD) is a major cause of morbidity and mortality and much international effort has been expended to detect risk factors, both heritable and environmental. Although there is a well established genetic basis for CAD, genome wide association studies (GWAS) have identified just 46 common loci, explaining only a small fraction (13%) of the predicted heritability of CAD, estimated by twin studies to be between 40 and 60%. This âAIJmissing heritabilityâĂİ, may be explained by diverse phenomenon including multiple common variants of very low effect size that may act via multiple causal risk factors for CAD and escape detection in sample sizes investigated to date, rare variants (MAF < 1%) of high effect size, gene  $\tilde{A}U$  gene ( $G\tilde{A}UG$ ) interactions, and gene  $\tilde{A}U$  environment ( $G \times G$ ) E) interactions. Previous efforts have tested the ability of a genetic risk score based on from 13 to 30 CAD-associated single nucleotide polymorphisms (SNPs) to predict CAD risk. Even this small number of risk alleles was shown to have significant predictive power and recently, to identify individuals who would benefit most from statin therapy to reduce LDL concentrations. However, improvements in genetic risk assessment are necessary and feasible given recent genetic advancements.

Purpose and Specific Objectives: This study hopes to develop an improved genetic risk score for coronary artery disease using a panel of independent risk loci. We address whether or not a panel of 202 independent SNPs with stepwise addition of cardiometabolic condition SNPs significantly predicts CAD.

Materials and Methods: 202 Independent SNPs were identified through GWAS and linear regression with multidimensional scaling in PLINK. The present study will use a stepwise logistic regression model with principal components and additional covariates. The independent variable will be a composite of genetic risk equal to a weighted sum of risk alleles with mean value imputation. The study will also compute NagelkerkeâĂŹs Psuedo-R2 as a proxy measure for goodness of fit of the model. Additionally, we will compute the receiver operator characteristic curve and calculate the area under the curve to determine model predictive accuracy. The net recombination index will also be calculated for each model. Accurate multiple correction will be performed with respect to the correlation matrix between tests. Additionally, the above

analysis will be repeated using different FDR thresholds using the R program PRSice.

**Results**: This study will result in several metrics describing the model's ability to predict CAD in a population. If the predictive ability of our score is meaningful, it will allow clinical researchers to diagnostically determine individual risk to CAD.

## Contents

Li	st of	Figure	es e	vi
Li	st of	Tables	3	viii
Li	st of	Acron	yms	xi
1	Intr	oducti	on	1
	1.1	Genet	ics of Coronary Artery Disease	2
	1.2	Genon	ne Wide Association Studies	4
		1.2.1	Primer on Genetics	4
		1.2.2	Sequencing	4
		1.2.3	Statistical Definition	4
		1.2.4	Multiple Comparisson Problem	6
	1.3	Polyge	enic Prediction of Complex Disease	6
	1.4	Polyge	enic Sliding Window Optimization	6
	1.5	Summ	ary	6
Bi	ibliog	graphy		7

# List of Figures

1.1	Progression of the formation of plaque causing Coronary Artery Disease (CAD).	
	Adapted from Gretch 2003	2
1.2	Fine mapping of the genomic interval on chromosome 9 associated with Coronary	
	Heart Disease. Adapted from Mcpherson et al 2006	3
1.3	Example of a Manhattan plot from a Genome wide Association Study (GWAS)	
	for CAD performed by Shunkert et al. 2011	5

# List of Tables

# List of Acronyms

CAD Coronary Artery Disease
PRS Polygenic Risk Score
oPRS Optimal Polygenic Risk Score
CARDIOGRAMC4D Coronary ARtery DIsease Genome wide Replication and Meta analysis plus The Coronary Artery Disease Genetics
GWAS Genome wide Association Study
GLC Global Lipids Consortium
GIANT The Genetic Investigation of ANthropometric Traits
BMI Body Mass Index
MI Myocardial Infarction
kb kilobase
<b>DNA</b> Deoxyribonucleic acid
<b>A</b> Adenine
C Cytosine4
<b>T</b> Thymine
<b>G</b> Guanine
logue energific genetic logation

CNV	copy number variant	4
InDel	l insertion/deletion	4
RNA	ribonucleic acid	4
SNP	single nucleotide polymorphism	_

#### Colophon

This document was typeset using the XeTeX typesetting system created by the Non-Roman Script Initiative and the memoir class created by Peter Wilson. The body text is set 10pt with Adobe Caslon Pro. Other fonts include Envy Code R, Optima Regular and. Most of the drawings are typeset using the TikZ/PGF packages by Till Tantau.

Introduction 1

As the efficiency and accuracy of rapid genome sequencing skyrockets, the potential for personalized therapies has made its way from science fiction to scientific reality. Using genetics to understand, diagnose, and eventually to predict illness is not a new idea; in recent years, however, technological ability and scientific understanding have advanced to such a point that researchers may predict risk for several diseases with reasonable confidence. Increasingly, variants in the human genome are being identified as being robustly linked to risk for complex illnesses such as heart disease [cite 9p21], obesity [cite fto], and schizophrenia [cite something]. However, much work remains to be done in order to create tools which may accurately predict individual disease risk from known and unknown genetic risk factors. In this thesis, we propose a novel extension to a well known methodology in order to better characterize disease risk from comorbid conditions using only summary statistics.

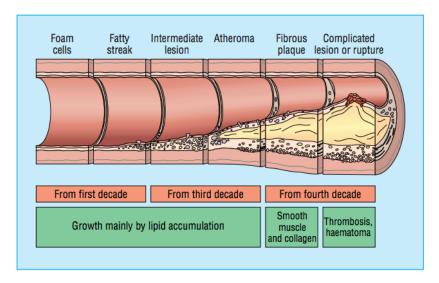
In brief, we present preliminary evidence for the use of Polygenic Risk Score (PRS)s in predicting CAD. We use recently published summary statistics from a GWAS conducted by the Coronary ARtery DIsease Genome wide Replication and Meta analysis plus The Coronary Artery Disease Genetics (CARDIOGRAMC4D) consortium alongside evidence gathered by the Global Lipids Consortium (GLC) and the The Genetic Investigation of ANthropometric Traits (GIANT) consortium for lipids and Body Mass Index (BMI) per say. We use this data alongside previously identified variants to construct first a simplistic PRS using only genome wide statistically signficant ( $P_{Bonferonni} < 0.05$  or  $q_{FDR} < 0.05$ ) variants, then expand our search to variants which may not be as robustly linked to phenotype. [Cite storey, BH, and dudbridge]. We use an empirical maximation approach and several strategies of mathematical optimization

in order to construct an *Optimal Polygenic Risk Score* (oPRS), then devise a novel technique for integrating information from co-morbid oPRS diseases in order to better predict CAD in four cohorts comprising approximately n = 12,000 individuals

#### 1.1 Genetics of Coronary Artery Disease

CAD occurs when the major blood vessels supplying the heart become diseased or damaged, often leading to severe complications such as *Myocardial Infarction* (MI) and death. [cite review articles] CAD is known to be a complex genetic disease with heritability estimated by twin studies between 40 and 60%. [mcPherson 2016, twin studies paper] Several important variants have been indentified which have been shown to robustly increase risk to CAD by altering lipid transporting pathways [cite LDLR], structural collegan bodies [TRIB 1??], and others factors.

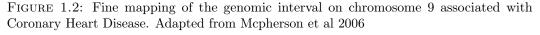
FIGURE 1.1: Progression of the formation of plaque causing CAD. Adapted from Gretch 2003.

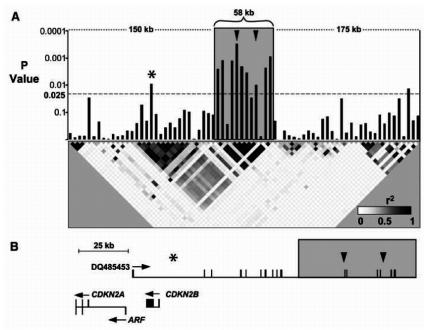


With heart disease and stroke the leading cause of perscription drug use in Canada as well as one of the leading causes of death and hospitalization [cite herat and stroke], the need to better understand, diagnose, and prevent this deadly disease is apparent. In order to better understand the need for improved statistical methodologies, it is important to understand the large body of previous attempts to characterize the genetic determinants of CAD

Despite some promising beginings, initial attempts to understand and explain CAD through genetics were largely unsuccessful. [CITE] The first variant to be successfully and robustly linked to risk for CAD was the 9p21.3 locus. Discovered by a team of researchers at the University of Ottawa Heart institute, the allele consists of a 58 kilobase (kb) region on chromosome 9 which was

shown to be associated with CAD in a population of 23,000 caucasion individuals. (McPherson and Tybjaerg-Hansen (2016))





This initial success began the era of the GWAS, explained in more detail in section 1.2. Researchers across the globe began frantically searching for more loci with the hope of understanding and predicting complex disease; in that goal, the GWAS has failed. (Visscher et al. (2012)) A number of important genetic markers for CAD have been discovered, but often in small familial cases or with very low effect sizes. [Cite] As the dust settles and the low hanging fruit have been picked, common variants have been shown to explain approximately 28% of the heritablity of CAD [cite majid], yet a large portion remains to be accounted for. This has become known as the problem of "missing heritability" of complex disease; common genetic variants explain a relatively small portion of the total estimated heritability of a disease, therefore researchers must resort to ever more obscure and complex methods to attempt to explain the complex interactions between genetic elements in the human genome. [cite review paper] From pathway analysis to partitioned heritability to all kinds of arcane statistical procedures, researchers from across the globe have tried their hardest to shrink this gap between our knowledge and accurate prediction and understanding of complex disease. To this end, we develop our own methodology incorporating multiple sources of information for the more accurate prediction of clinical end points.

#### 1.2 Genome Wide Association Studies

In order to properly introduce the model, however, the basic underpinnings must be explored and explained. Genome wide association studies seek to indetify associations between individual genotypes and disease phenotypes in a hypothesis free manner. In this section, the statistical model required to understand GWAS is presented and explored.

#### 1.2.1 Primer on Genetics

Deoxyribonucleic acid (DNA) is a double helical molecule which encodes the genetic blueprints for the construction of proteins and other materials that make up every known living organism. DNA is composed of three parts: a negatively charged phosphate group, a five carbon sugar deoxyribose, and (usually) one of four nitrogen bases. It is these bases, Adenine (A), Cytosine (C), Thymine (T), and Guanine (G) and their combinations which are under investigation in a GWAS. The specific combinations of these four bases in a specific genetic location (locus) determine the product produced by the DNA, and even a small change in this order can have large ramifications on the overall health, survival, and proper function of the organism.

#### 1.2.2 Sequencing

DNA sequencing is the process of ascertaining a particular individual's genotype by means of chemical identification of the bases present at predefined sites. [cite] These sites, whether they be a change in a single base called a *single nucleotide polymorphism* (SNP), a variation in the number of tandem repeats of a small sequence named a *copy number variant* (CNV) or an *insertion/deletion* (InDel) of a sequence, may alter amino acid sequence, affect regulatory regions, or impact regulatory *ribonucleic acid* (RNA) sequences.

**Definition 1.2.1 (Allele)** A specific form or subtype of a genetic locus. This could be one or more individual variations or a combination thereof.

**Remark 1** Allele frequency is the frequency at which a particular allele occurs in the population. I.e. for locus A having n different alleles, the true population allele frequency of allele  $freq A_m \equiv \frac{A_m}{\sum_{i=1}^n A_i}$ , which is estimated in a sample population with a biased ratio estimator  $freq \hat{A}_m \equiv \frac{\hat{A}_m}{\sum_{i=1}^n \hat{A}_i}$ 

#### 1.2.3 Statistical Definition

Consider a simple case control population where 1 defines case and 0 defines control. Define  $\underline{\mathbf{Y}}$  as an *n*-vector where *n* denotes the number of individuals in a population and  $\underline{\mathbf{Y}}_i$  gives the

individual's diesease staet. Additionally define G as an  $m \times n$  matrix where m is the number of informative genotypic sites available with  $\mathbf{G}_{ij}$  being the "state" (allele number) present at site  $j, 1 \le i \le m, i \in \mathbb{Z}^+$  in individual  $i, 1 \le i \le n, i \in \mathbb{Z}^+$ .

$$\underline{\mathbf{Y}} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}$$
 $\mathbf{G} = \begin{bmatrix} G_{1,1} & \dots & G_{1,n} \\ \vdots & \ddots & \vdots \\ G_{m,1} & \dots & G_{m,n} \end{bmatrix}$ 

In an additive genetic model, we define the phenotype  $\underline{\mathbf{Y}}$  as a linear combination of  $\mathbf{G}$  weighted by a vector of  $\underline{\beta}$  coefficient vectors estimated by regression analysis and  $\underline{\epsilon}$  vector of errors. Express  $\underline{\mathbf{Y}}$  such that

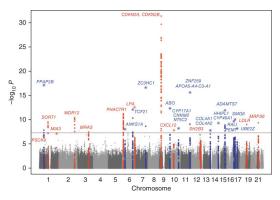
$$\underline{\mathbf{Y}} = \underline{\beta}' \mathbf{G} + \underline{\epsilon} = \left( \sum_{i=1}^{m} \beta_i \mathbf{G}_{i,n} + \epsilon_n \right)'$$

 $\underline{\beta}$  and  $\underline{\epsilon}$  are approximated optimally by  $\underline{\hat{\beta}}$  and  $\underline{\hat{E}}$  in practice.

The purpose of a GWAS is not only to estimate these genetic effects  $\underline{\beta}$  by  $\underline{\hat{\beta}}$  but also to estimate their significance of association with phenotype vector  $\underline{\mathbf{Y}}$  through a  $\chi^2$  test and corresponding test statistic m-vector  $\hat{\chi}^2$ . The degrees of freedom of this test statistic will vary between methods and models, and so will be left as futher reading.

By approximating  $\underline{\chi^2}$  with  $\underline{\hat{\chi^2}}$  and computing the corresponding P values, reserchers are able to identify and quantify the effects of variants significantly (P < 0.05) associated with the phenotype. These results can be summarized in a Manhattan plot, named after the city of Manhattan with it's high rise buildings towering over the scenery. The x axis of this plot is the genomic location (usually coloured by chromosome number) while the y axis is the  $\log_{10}$  of the P value of association derived from  $\hat{\chi^2}$ .

Figure 1.3: Example of a Manhattan plot from a GWAS for CAD performed by Shunkert et al. 2011



#### 1.2.4 Multiple Comparisson Problem

In such a set up, where m may be in the millions and the threshold of significance is set to  $P = \alpha = 0.05$ , we encounter a canonical issue in statistical inference. Recall that P is the probability of observing a  $\chi^2$  statistic as large or larger than a specific  $\chi^2_m$  assuming  $H_0$  of no association is correct and  $\alpha$  is the threshold at which a significant effect is declared.

### 1.3 Polygenic Prediction of Complex Disease

Refering to the defintions proposed in the previous section and recalling that

### 1.4 Polygenic Sliding Window Optimization

### 1.5 Summary

## Bibliography

Ruth McPherson and Anne Tybjaerg-Hansen. Genetics of Coronary Artery Disease. *Circulation Research*, 118(4):564-578, feb 2016. ISSN 0009-7330. doi: 10.1161/CIRCRESAHA.115.306566. URL http://circres.ahajournals.org/content/118/4/564.abstract.

Peter M Visscher, Matthew A Brown, Mark I McCarthy, and Jian Yang. Five years of GWAS discovery. American journal of human genetics, 90(1):7-24, jan 2012. ISSN 1537-6605. doi: 10.1016/j.ajhg.2011.11.029. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3257326{&}tool=pmcentrez{&}rendertype=abstract.