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

May 2016

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 âĂIJmissing 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 ÃŮ gene (GÃŮG) interactions, and gene ÃŮ environment (G $\times$ 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

List of Figures

List of Tables

List of Acronyms

CAD Coronary Artery Disease vi

PRS Polygenic Risk Score 1

oPRS Optimal Polygenic Risk Score 2

CARDIOGRAMC4D Coronary ARtery DIsease Genome wide Replication and Meta analysis plus The Coronary Artery Disease Genetics 1

GWAS Genome wide Association Study 1

GLC Global Lipids Consortium 1

GIANT The Genetic Investigation of ANthropometric Traits 1

BMI Body Mass Index 1

MI Myocardial Infarction 2

kb kilobase 2

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.

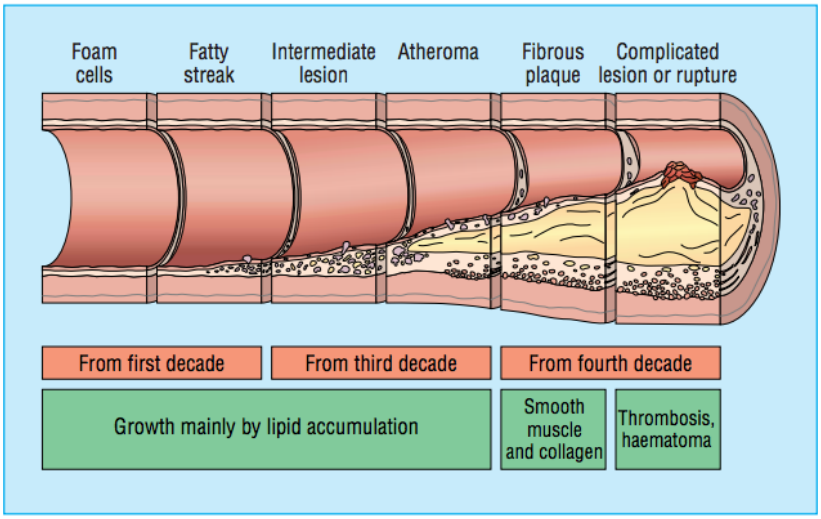
In brief, we present preliminary evidence for the use of *Polygenic Risk Score* (PRS)s in predicting CAD. We use recently published summary statistics from a *Genome wide Association Study* (GWAS) conducted by the *Coronary ARtery DIsease Genome wide Replication and Meta analysis plus The Coronary Artery 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.

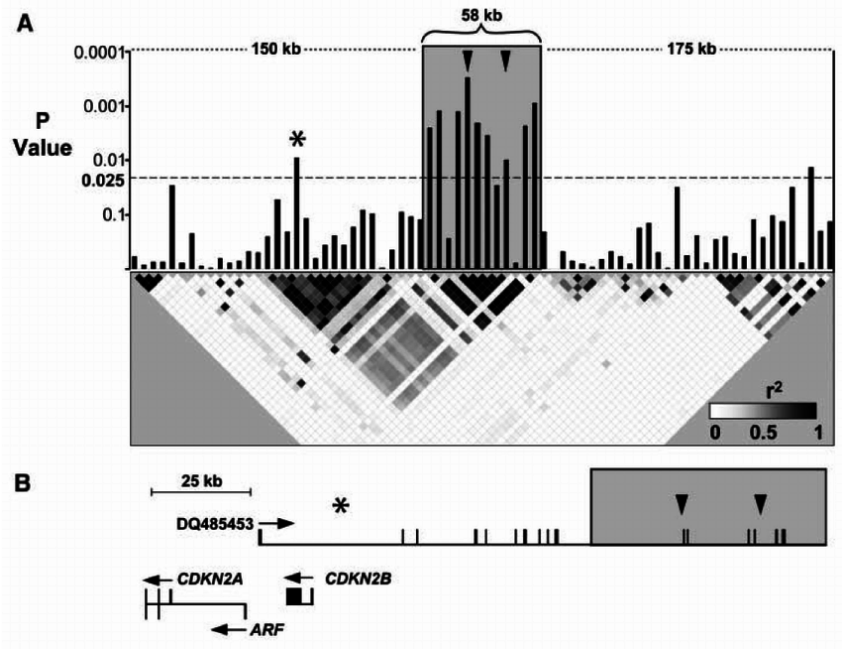


With heart disease and stroke the leading cause of perscription drug use in Canada as well as one of the leading causes of death and hospitalization [cite herat and stroke], the need to better understand, diagnose, and prevent this deadly disease is apparent. In order to better understand the need for 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 succesfully and robustly linked to risk for CAD was the 9p21.3 locus. Discovered by a team of researchers at the Univeristy 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))

FIGURE 1.2: Fine mapping of the genomic interval on chromosome 9 associated with Coronary Heart Disease. Adapted from Mcpherson et al 2006



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))

1.2 Genome Wide Association Studies

1.3 Polygenic Prediction of Complex Disease

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`.