GENETICS OF CORONARY ARTERY DISEASE

GENETIC DETERMINANTS OF RARE CODING VARIANTS ON THE DEVELOPMENT OF EARLY-ONSET CORONARY ARTERY DISEASE

BY RICKY LALI, B.Sc. (Hon.)

A Thesis Submitted to the School of Graduate Studies in Partial Fulfillment

of the Requirements for the Degree Master of Science

MASTER OF SCIENCE (2017) McMaster University

(Biochemistry and Biomedical Sciences) Hamilton, Ontario

TITLE: Genetic Determinants of Rare Coding Variants on the Development of Early-Onset Coronary Artery Disease

AUTHOR: Ricky Lali, B.Sc. (Hon.) (McMaster University, 2012 & 2014)

SUPERVISOR: Guillaume Paré, MD MSc

NUMBER OF PAGES: xxiv, 186

ABSTRACT

Background: Coronary Artery Disease (CAD) represents the leading cause of mortality and morbidity worldwide despite declines in the prevalence of environmental risk factors. This trend has drawn attention to the risk conferred by genetic variation. Twin and linkage studies demonstrate a profound hereditary risk for CAD, especially in young individuals. Rare genetic variants conferring high risk for extreme disease phenotypes can provide invaluable insight into novel mechanisms underlying CAD development.

Methods: Whole exome sequencing was performed to characterize rare protein-altering variants in 52 early-onset CAD (EOCAD) patients encompassing the DECODE study. The enrichment of Mendelian dyslipidemias in EOCAD was assessed through interrogation of pathogenic mutations among known lipid genes. The identification of novel genetic CAD associations was conducted through case-only and case-control approaches across all protein-coding genes using rare variant burden and variance component tests. Lastly, beta coefficients for significant risk genes from the European population in the Early-onset Myocardial Infarction (EOMI) cohort (N=552) were used to construct calibrated, single-sample rare variant gene scores (RVGS) in DECODE Europeans (N=39) and a local European CAD-free cohort (N=77).

Results: A 20-fold enrichment of Familial hypercholesterolemia mutation carriers was detected in EOCAD cases compared to CAD-free controls (P=0.005). Association analysis using EOMI Europeans revealed exome-wide and nominal significance for two known CAD/MI genes: CELSR2 ($P=1.1x10^{-17}$) and APOA5 (P=0.001). DECODE association revealed exome-wide and nominal significance for genes involved in endothelial integrity

and immune cell activity. RVGS based upon beta coefficients of significant CAD/MI risk genes were significantly increased in DECODE (z-score=1.84; p=0.03) and insignificantly decreased among CAD-free individuals (z-score=-1.61; p=0.053).

Conclusion: Rare variants play a pivotal role in the development early CAD through Mendelian and polygenic mechanisms. Construction of RVGS that are calibrated against population and technical biases can facilitate discovery of single-sample and cohort-based associations beyond what is detectable using standard methods.

ACKNOWLEDGMENTS

I would like to begin this section by stating that the work leading to thesis has been among the most rewarding experiences I have encountered. This work has not only provided me with a strong knowledge-base in genetics and biostatistics, but has also gifted me the opportunity to work alongside truly amazing individuals from whom I have learnt a great deal and forged lasting friendships.

I would foremost like to give unwavering thanks to my supervisor and mentor, Dr. Guillaume Paré. The time and care you devote towards ensuring each of your students grasps a genuine understanding of their work is nothing short of remarkable and is something I wish to emulate. Thank you for devoting countless hours in teaching me the fundamentals of good research practice and for providing me an opportunity to attain skillsets that will prove invaluable for achieving success in my future endeavors. This work would not have been possible without your guidance. I wish to also pay thanks to my committee members, Dr. Eva Szabo and Dr. Madhu Natarajan for their wonderful support in making this project a success and for their leadership in overseeing the different arms of this study.

I would next like give a special thanks to Michael Chong, whose friendship and guidance were pivotal to the success of this work. I wish to also extend thanks to the entire Genetics and Molecular Epidemiology Laboratory (GMEL) team for their wonderful support over the last 2.5 years. Thank you to Taylor MacIsaac, Shana Hayter, Amanda Hodge, and Reina Ditta for performing and managing all sequencing and genotyping operations. Thank you to Sébastien Thériault and Marie Pigeyre for providing me with necessary clinical data. Thank you to Michael Chong, Jenny Sjaarda, Kripa Raman, Shihong Mao, and Pedrum Mohammadi-Shemirani for assisting with all my statistical and bioinformatics inquiries.

I would also like to extend a very special thank you to Chelsia Baral, who never failed to bring a smile to my face or provide me with an escape from the rigors of life.

Lastly, I would like to pay a very special thanks to my loving family: my sister, Rupinderjeet Singh; my brother, Ravinderjeet Lali; my mother Kusum Lali; and my father, Kanwaljit Lali. I would not have been able to complete this work if not for your unconditional support and encouragement. You all epitomize the spectacle of perseverance and diligence. I would like to close by dedicating this work to my father. It goes without saying that you have made all my aspirations a reality. I will likely never accomplish the feats you have, but take immense solace in consistently striving to conduct myself in a manner you would see fit. I will forever be proud to be your son.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1 – General Introduction

CHAPTER 2 – DECODE: Project Proposal for an Early-Onset Coronary Artery Disease Cohort

CHAPTER 3 – Whole Exome Quality Control: Understanding Patterns of Genetic Variation

CHAPTER 4 – Assessing the Prevalence of Mendelian Dyslipidemias in an Early CAD Population and Considerations for Clinical Intervention

CHAPTER 5 – Rare Variant Association: Leveraging External and Internal Control Datasets for Gene Discovery

CHAPTER 6 – 'N of 1' Benchmarking for the Calibration of Individual Sequences to Big Data: A Novel Methodology to Facilitate Construction of Rare Variant Gene Risk Scores

SUPPLEMENTAL TABLES

LIST OF FIGURES

CHAPTER 1 – General Introduction

CHAPTER 2 – DECODE: Project Proposal for an Early-Onset Coronary Artery Disease Cohort

CHAPTER 3 – Whole Exome Quality Control: Understanding Patterns of Genetic Variation

CHAPTER 4 – Assessing the Prevalence of Mendelian Dyslipidemias in an Early CAD Population and Considerations for Clinical Intervention

CHAPTER 5 – Rare Variant Association: Leveraging External and Internal Control Datasets for Gene Discovery

CHAPTER 6 – 'N of 1' Benchmarking for the Calibration of Individual Sequences to Big Data: A Novel Methodology to Facilitate Construction of Rare Variant Gene Risk Scores

LIST OF ABBREVIATIONS

1KGP3	1000 Genomes Project Phase 3
AAF	Alternate Allele Frequency
ABCA1	ATP-Binding Cassette Family A Member 1
ABCC6	ATP-Binding Cassette Family C Member 6
ABCG5/8	ATP-Binding Cassette Family G Member 5/8
ACC	American College of Cardiology
ACMG	American College of Medical Genetics
ACS	Acute Coronary Syndromes
AFR	Africans
ALPI	Intestinal Alkaline Phosphatase
AMP	Association for Molecular Pathology
AMR	Latin Americans
ANGPTL4	Angiopoietin like-4
APOA5/V	Apolipoprotein A5/V
APOB	Apolipoprotein-B100
APOC2	Apolipoprotein-C2
APOC3	Apolipoprotein-C3
APOE	Apolipoprotein-E
BAM	Binary Alignment
bFGF	basic Fibroblast Growth Factor
BMI	Body Mass Index
BrDu	Bromodeoxyuridine
BTLN3	Butyrophilin like 3
CAD	Coronary Artery Disease
CADD	Combined Allelic Dependent Depletion
CARDIoGRAMplusC4D	Coronary ARtery DIsease Genome wide Replication and Meta-analysis plus The Coronary Artery Disease

CBS	Cystathionine Beta-Synthase
CCL3/4	Chemokine c-c Motif Ligand 3/4
CCN	Cardiac Care Network
CEACAM1	Carcinoembryonic antigen-related Cell Adhesion Molecule 1
CELSR2	Cadherin EGF LAG seven-pass G-type receptor 2
CES	Clinical Exome Sequencing
CF	Correction Factor
CHARGE	Cohorts for Heart and Aging Research in Genomic Epidemiology
CI	Confidence Interval
CKD	Chronic Kidney Disease
CLEC4D	C-Type Lectin Domain Family 4 Member D
CMAC	Cumulative Minor Allele Count
CMAF	Cumulative Minor Allele Frequency
CML	Candidate Mapping Location
CRCTL	Clinical Research and Clinical Trial
cTnI/T	Cardiac Troponin I/T
CVCD	Common Variant-Common Disease Hypothesis
CVD	Cardiovascular Disease
DALYS	Disability-Adjusted life years
dbGaP	The Database of Genotypes and Phenotypes
dbSNP	The Single Nucleotide Polymorphism Database
DHX34	DExH-Box Helicase 34
DP	Depth of Coverage
EAS	East Asians
EC	Endothelial Cell
ECE2	Endothelin-Converting Enzyme 2
ECM	Extracellular Matrix
EOCAD	Early-Onset Coronary Artery Disease

EOMI	Early-Onset Myocardial Infarction
EPS	Extreme Phenotype Sampling
ExAC	Exome Aggregation Consortium
FCH	Familial Combined Hyperlipidemia
FDR	False Discovery Rate
FH	Familial Hypercholesterolemia
FIN	Finnish Europeans
FS	Frameshift
GATK	Genome Analysis Toolkit
GIAB	Genome In A Bottle
GLGC	Global Lipids Genetic Consortium
gnomAD	Genome Aggregation Database
GPIHBP1	Glycosylphosphatidylinositol Anchored High Density Lipoprotein Binding Protein 1
GRS	Gene Risk Score
GWAS	Genome-Wide Association Studies
HARPS	Heart Attack Risk in Puget Sound
HDL-C	High-Density Lipoprotein Cholesterol
HET	Heterozygous
HGNC	HUGO Gene Nomenclature Committee
HGVS	Human Genome Variation Society
HOM	Homozygous
HRC	Haplotype Reference Consortium
HRUN	Homopolymer Run
HSD3B7	Hydroxy-delta-5-Steroid Dehydrogenase, 3 Beta- Steroid Delta-Isomerase 7
ICAM-1	Intercellular Adhesion Molecule 1
IL7R	Interleukin 7 Receptor
INDEL	Insertion/Deletion

IQR	Inter Quartile Range
iPSC	Induced Pleuripotent Stem Cell
ISP	Ion Sphere Particle
k_0	Kinship Coefficient
LD	Linkage Disequilibrium
LDL-C	Low-Density Lipoprotein Cholesterol
LDLR	Low-Density Lipoprotein Receptor
LDLRAP1	Low-Density Lipoprotein Receptor Associated Protein 1
LOF	Loss-of-Function
LOX1	Lectin-like Oxidized Receptor 1
LPL	Lipoprotein Lipase
LPS	Lipopolysaccharides
MAF	Minor Allele Frequency
M-CAP	Mendelian Clinically Applicable Pathogenicity
MCP-1	Monocyte Chemo-Attractant Protein 1
mCSF	Macrophage-Colony Stimulating Factor
$MgCl_2$	Magnesium Chloride
MGH-PCAD	Massachusetts General Hospital Premature Coronary Artery Disease Study
MI	Myocardial Infarction
MIGen	Myocardial Infarction Genetics
MLLD	Relative read quality
MMP	Matrix-Metalloproteinase
MMP9	Matrix Metalloproteinase 9
MTHFR	Methylenetetrahydrofolate Reductase
MTMR9	Myotubularin Related Protein 9
ncRNA	Non-Coding RNA
NFE	Non-Finnish Europeans

NF-kB	Nuclear Factor Kappa-light-chain-enhancer of activated B cells
NFS	Non-Frameshift
NHLBI GO ESP6500	National Heart, Lung and Blood Institute Grand Opportunity Exome Sequencing Project v.6500
NPC1L1	Niemann–Pick C1-like 1
NS	Nonsynonymous
NSTEACS	Non-ST-Elevated Acute Coronary Syndromes
NSTEMI	Non-ST-Elevated Myocardial Infarction
OCT4	Octamer-binding Transcription Factor 4
OR	Odds Ratio
oxLDL-C	Oxidized Low-Density-Lipoprotein Cholesterol
PAR	Population Attributable Risk
PCR	Polymerase Chain Reaction
PCSK9	Proprotein Convertase Subtilisin/Kexin type 9
PDGF	Platelet Derived Growth Factor
PNPLA5	Patatin-like phospholipase domain-containing 5
Polyphen2	Polymorphism Phenotype v2
QC	Quality Control
qPCR	Quantitative Polymerase Chain Reaction
RCA	Right Coronary Artery
RefSeq	NCBI Reference Sequence Database
RR	Relative Risk
RVAS	Rare Variant Association Analysis
RVB	Rare Variant Burden Score (aggregate)
RVGS	Rare Variant Gene Score
S	Synonymous
SAS	South-East Asians
SCARA1/2	Scavenger Receptor Class A Member I/II

SCARB1	Scavenger Receptor Class B Member I
SD	Standard Deviation
SHAPEIT	Segmented Haplotype Estimation and Imputation Tool
SIFT	Sorting Intolerant From Tolerant
SKAT	Sequence Kernal Association Test
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variant
SOX2	SRY-box 2
STAP1	Signal Transducing Atapter Protein 1
STB	Strand Bias
STEMI	ST-elevated Myocardial Infarction
TG	Triglycerides
TH1	Type 1 T helper
TMAP	Torrent Mapping Alignment Program
TRL	Triglyceride Rich Lipoprotein
TRIUMPH	Translational Research Investigating Underlying Disparities in Myocardial Infarction Patients' Health Status
TVC	Torrent Variant Caller
UA	Unstable Angina
UTR3	3-prime Untranslated Region
UTR5	5-prime Untranslated Region
VCAM-1	Vascular-Cell Adhesion Molecule 1
VCF	Variant Call File
VLDL	Very Low-Density Lipoprotein
VSMC	Vascular Smooth Muscle Cell
VUS	Variants of Uncertain Significance
WDLV	Western Database of Lipid Variants
wMACF	Weighted Minor Allele Carrier Frequency
wMAF	Weighted Minor Allele Frequency

CHAPTER 1 – General Introduction

1.1 Coronary Artery Disease: Environment versus Genetics

Cardiovascular diseases (CVDs) represent the number one cause of death worldwide despite advancements in both preventative and therapeutic care [1–3]. A significant proportion of CVDs (~42%) are attributed to Coronary Artery Disease (CAD), which is an inflammatory-metabolic disorder characterized by the formation of atherosclerotic plaques within the coronary arteries [4]. These plaques directly contribute to the development of arterial stenosis which can result in both acute and chronic clinical complications such as myocardial infarctions and heart failure, respectively. Due to the severity of these clinical outcomes, CAD was recently established as the leading cause of global disease burden as measured by disability-adjusted life years (DALYS), which renders CAD is the leading cause of healthy years lost [4]. Therefore, there is a strong initiative to establish diagnostic tests that can predict an individual's risk for developing CAD at an early stage in order to facilitate timely medical intervention and better patient outcomes. Genetic testing offers a robust approach to predict onset and severity of CAD based on determining an individual's genotype at risk loci. However, this remains a difficult endeavor as CAD is a complex disease driven by multiple genetic and environmental variables [4–6].

The environmental risk factors predisposing individuals to CAD have been well established, and commendable efforts have been made by health policy advocates to reduce their exposure to the general population. However, risk factors such as smoking, cholesterol, diabetes, hypertension, and abdominal obesity are still key modifiable contributors to CAD [7]. INTERHEART, a global case-control initiative sought to identify the magnitude of effect (odds ratio (OR)) conferred by these environmental factors on CAD

risk and determine the proportion of disease incidence that would be prevented if exposure was eliminated (i.e. the population attributable risk (PAR)) [7]. The 5 bona-fide CAD environmental risk factors collectively accounted for a PAR of 80% and an OR of 68, which explains the majority of CAD risk (Table 1). However, seminal studies have methodically established that first-degree family history (that is parental CVD risk conferred on offspring) independently predicts CVD events in offspring [8]. For instance, the Framingham Offspring Study conducted prospective follow-up in the offspring (N=2302) of participants in the original Framingham cohort who had either suffered a CVD event or were CVD free [8]. Follow-up of offspring revealed that individuals with one or more parent afflicted with CVD were significantly more likely to manifest a cardiovascular event than individuals with parents that had not suffered a CVD (OR 3.0 male and 2.6 female) [8].

The magnitude of risk for offspring CVD was slightly attenuated after accounting for the 5 environmental risk factors investigated in the INTERHEART study, but remained significant (OR 1.5 male and 1.1 female) [8]. A substantially stronger effect was observed when the analysis was restricted to offspring with *both* parents affected by *premature* CVD (i.e. father $\leq$ 55 years and mother $\leq$ 65 years) even after environmental risk-factor adjustment (OR 2.4 male and 2.8 female) [8].

These findings provide robust evidence that CAD manifestation (among other CVDs) is dependent on one's genetic composition, independent of environmental risk factors. Also, the results present a strong rationale for delineating genetic contributors to disease onset, especially among individuals that develop premature disease as they have not accumulated exposure to environmental risk factors.

Table 1.1: Comparison of the magnitude of effect conferred by environmental versus heritable CAD risk factors.

Risk factor	**OR**	**PAR %**
Smoking	2.78	35.7
Apolipoprotein-B/Apolipoprotein-A1	3.25	49.2
Diabetes	2.37	9.9
Hypertension	1.91	17.9
Abdominal obesity	1.12	20.1
All of above	**68.5**	**80**
Parental CVD - 1 or both parents (multivariate adjusted)	1.5 - male	23.1 †
	1.1 - female	5.7 †
Parental CVD – Both parents (multivariate adjusted)	1.8 - male	32.5 †
	1.0 - female	0 †
Premature parental CVD – 1 or both parents (multivariate adjusted)	2.0 - male	37.6 †
	1.7 - female	29.6 †
Premature parental CVD – Both parents (multivariate adjusted)	**2.4 - male**	**45.7** †
	2.8 - female	**52** †

All factors were found to be statistically significant ($P<0.05$)
† PAR for heritable risk factors were estimated from OR and in-study CVD prevalence

1.2 Twin studies and Mendelian disease: Establishing CAD heritability

The genetic basis of CAD has been established through twin studies and cascade analysis of families with significant disease history. Twin studies demonstrate the risk of developing disease in monozygotic and dizygotic twin pairs, which share 100% and 50%

genomic identity, respectively. A benefit of twin studies is that they inherently control for extraneous variables that may otherwise contribute to disease manifestation (e.g. age and diet). As a result, differences observed in disease risk between monozygotic and dizygotic twins can largely be attributed to a genetic source. Marenburg *et al*. conducted a large longitudinal study of 10 000 monozygotic and dizygotic twin pairs and observed that monozygotic twins have a two-fold increase in relative risk (RR) over dizygotic twins for death by early onset CAD if their co-twins had previously died from the disorder (RR = 8.1 vs. 3.8) [6,9]. Interestingly, the RR steadily increased as age at death by CAD decreased [9], further suggesting that a genetic contribution to disease onset may be more prevalent in younger people.

In addition to twin studies, familial cascade testing has provided evidence supporting a hereditary mechanism for disease onset. These studies are typically conducted by surveying families with significant disease history and identifying whether putative disease-causing mutations at candidate loci co-segregate with the phenotype of interest. Profound co-segregation can affirm a causal relationship between gene and disease and can also indicate on whether the mode of inheritance is dominant or recessive. However, most pedigrees are confounded by extraneous variables that can impede a clear genotype-phenotype relationship from being established. Common confounders include 1) polygenic modes of inheritance (i.e. multiple genes additively contributing to disease risk), 2) incomplete penetrance, 3) presence of causal risk factors within affected individuals, and 4) clinical mis-diagnoses of disease phenotype to due sub-clinical manifestations of disease.

Mendelian disorders are typically free of these confounders as they involve rare, highly penetrant mutations that confer their effect through a single gene with a defined pattern of inheritance. As such, Mendelian diseases that manifest with CAD as a primary phenotype have provided invaluable evidence to support a genetic contribution to disease onset. The most prominent Mendelian diseases that manifest with CAD are monogenic dyslipidemias, which result in severely perturbed lipoprotein metabolism and are sufficient towards promoting formation of atherosclerotic lesions. Well-established Mendelian disorders that affect lipoprotein metabolism include Familial hypercholesterolemia, Familial combined hyperlipidemia, Sitosterolemia, and Autosomal dominant coronary artery disease 2 [10]. These disorders elicit their affects through either an autosomal dominant or autosomal recessive mode of inheritance, depending on the affected gene (Table 1.2) [10]. An example of CAD manifesting in a Mendelian pattern is depicted in Figure 1.1 which shows pedigrees of two British families affected with FH characterized by individuals with history of elevated LDL-C and CAD/MI in multiple generations [11]. Individuals shown as half-shaded are heterozygote carriers of the p.Asp374Tyr gain-of-function mutation in (PCSK9), which had been shown to confer considerable risk for premature CAD. In both pedigrees, carrier status co-segregates completely with both CAD/MI and severe dyslipidemia, thus establishing a causal genotype-phenotype relationship assuming no other candidate mutations were identified. Absence of CAD/MI in some p.Asp374Tyr carriers can be attributed to early therapeutic intervention that more than likely prevented a clinical outcome.

Both familial cascade analysis and twin studies have benchmarked the genetic foundation of CAD development. By virtue of their study designs, these analyses are well structured to determine what proportion of phenotypic variability can be attributed to genetic or environmental causes. The heritability estimate corresponds to the ratio of additive genetic variance to total phenotypic variance for a given trait and provides a quantitative measure to assess the strength of genetic contribution to disease development. Both the early twin and cascade studies have established a heritably of 40-50% for CAD [2,5,10], meaning that approximately half the phenotypic variability can be ascribed to genetic factors. Interestingly, it has been shown that CAD heritability is approximately 63% for early-onset cases, underscoring the notion that premature CAD is largely explained by an individual's genetic composition [10].

Table 1.2: Summary data of Mendelian diseases with CAD as a hallmark phenotype (adapted from Stitziel *et al.* 2014) [10].

Disease category	Condition	Gene(s) Involved	Mode of Inheritance	Major cardiovascular clinical outcomes
Mendelian dyslipidemias	Autosomal dominant familial hypercholesterolemia	LDLR, APOB, PCSK9, STAP1	Autosomal dominant	>99th percentile plasma LDL-C; premature CAD
	Autosomal recessive familial hypercholesterolemia	LDLRAP1	Autosomal recessive	>99th percentile plasma LDL-C; premature CAD
	Familial combined hyperlipidemia	LPL	Autosomal dominant	Elevated plasma triglycerides, LDL-C and APOB; small, dense LDL-C; premature CAD
	Sitosterolemia	ABCG5, ABCG8	Autosomal recessive	Elevated campesterol and sitosterol; premature CAD
	Autosomal dominant coronary artery disease 2	LRP6	Autosomal dominant	Metabolic syndrome; premature CAD
Mendelian vasculaopathies	Homocystinuria	CBS, MTHFR	Autosomal recessive	Elevated homocysteine; premature CAD
	Pseudoxanthoma elasticum	ABCC6	Autosomal recessive	Enhanced coronary stenosis and calcification; premature CAD

LDLR = Low-density lipoprotein receptor; APOB = Apolipoprotein B-100; PCSK9 = Proprotein convertase subtilisin/kexin type 9; STAP1 = Signal Transducing Adaptor Family Member 1; LDLRAP1 = Low Density Lipoprotein Receptor Adaptor Protein 1; LPL = Lipoprotein lipase; ABCG5/8 = ATP binding cassette transporter family G type 5/8; LRP6 = LDL Receptor Related Protein 6; CBS = Cystathionine beta-synthase; MTHFR = Methylenetetrahydrofolate reductase; ABCC6 = ATP binding cassette transporter family C type 6

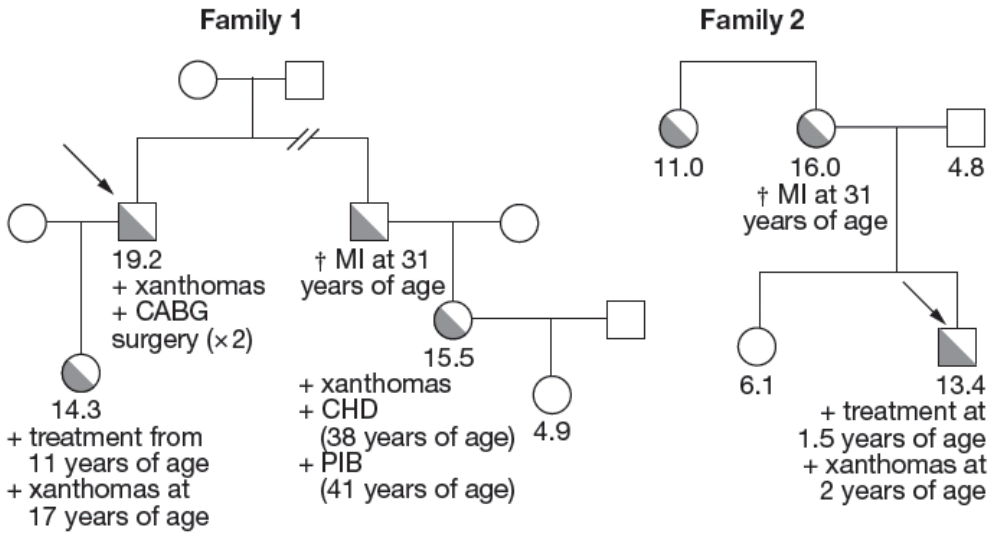


Figure 1.1: Co-segregation of dyslipidemia and early CAD/MI with carrier status of the Arg474Tyr mutation in PCSK9 for two British families. Mutation carriers are shown as half shaded squares/circles and LDL-C values are shown immediately below individual symbols followed by pertinent phenotypic information. Squares and circles represent males and females, respectively. Arrows indicate the index cases in each family.

1.3 Genome-wide association studies and the common variant-common disease hypothesis

Genetic variants that confer risk for CAD development can be identified through Genome Wide Association Studies (GWAS), which are case-control analyses that seek to elucidate statistically significant differences in the abundance of known genetic variants between cases individuals afflicted with disease phenotype (cases) and disease-free individuals (controls) across the entire genome [5,12]. The variants analyzed in these studies occur at an intermediate to high frequency (i.e. Minor Allele Frequency (MAF) > 5%) in the study populations [5,12]. These *common* variants are preferentially queried in most case-control association studies for two major reasons: (1) rarer variants typically do not meet the statistical threshold necessary to classify them as significant (especially while using

single-locus association models) and (2) common variants greatly contribute to disease prevalence despite typically having small effects on disease phenotype [13]. Therefore, information from multiple common variants would have to be genotyped in a given individual before their risk stratification could be improved, which is the underlying motivation for the development of gene risk scores (GRS).

The Coronary ARtery DIsease Genome wide Replication and Meta-analysis plus The Coronary Artery Disease (CARDIoGRAMplusC4D) consortium represents the largest aggregated dataset of genotyped CAD cases + controls (N=185000) and has provided invaluable insight into the genetic architecture of CAD. CARDIoGRAMplusC4D was conducted predominantly within individuals of European ancestry and has identified 58 loci associated with CAD at genome-wide significance under either additive or recessive models of inheritance [5]. However, a total of 129 loci were found to demonstrate either significant or suggestive associations with CAD at a false discovery rate of 5% (129 FDR). The MAFs among the genome-wide significant loci were largely common (median MAF 0.22) and conferred only modest increases in CAD risk (median OR 1.07 for effect alleles conferring risk and OR 0.93 for effect alleles conferring protection). Collectively, the variants driving the 129 FDR association accounted for approximately 22% of CAD heritability, which is estimated to be at 40-50% based on seminal cascade and twin studies discussed earlier [5]. Therefore, a large proportion of the heritability remains missing and may have to be accounted for through the discovery of additional common variants of low to modest effect or by low frequency (1% < MAF < 5%) and rare (MAF < 1%) variants conferring larger effects on disease risk. Overall, the CARDIoGRAMplusC4D meta-

analysis queried ~9.4 million variants for association with CAD. Over 90% of these variants fell within either intergenic or intronic regions (Figure 1.2) which will preferentially lead to the discovery of alleles that mediate CAD risk through regulatory effects such as the altering of gene expression through disruption of transcription factor binding sites or modification of epigenetic hotspots. In contrast, only ~0.6% (58160 / ~9 million total variant sites) of queried variants were located within the coding sequence or in canonical splice donor/acceptor sites. Considering that coding regions comprise 1.5-2% of the genome, this proportion substantially below what would be expected. Therefore, the risk elicited by higher impact variants that can alter protein structure remains largely unknown and must be accounted for by exome sequencing strategies.

The loci found to reach genome wide significance were classified into gene ontologies known to be involved in the pathogenesis of CAD (section 1.8) such as lipid metabolism, endothelial integrity, and haemostasis (Figure 1.2) which emphasizes the essential role of these biological processes in conferring lifelong risk for CAD. However, the mechanisms by which many of the genome-wide significant loci confer CAD risk remain unknown. Conducting a systematic analysis to identify enrichment of rare variants in nearby genes can potentially lead to the discovery biologically relevant candidates.

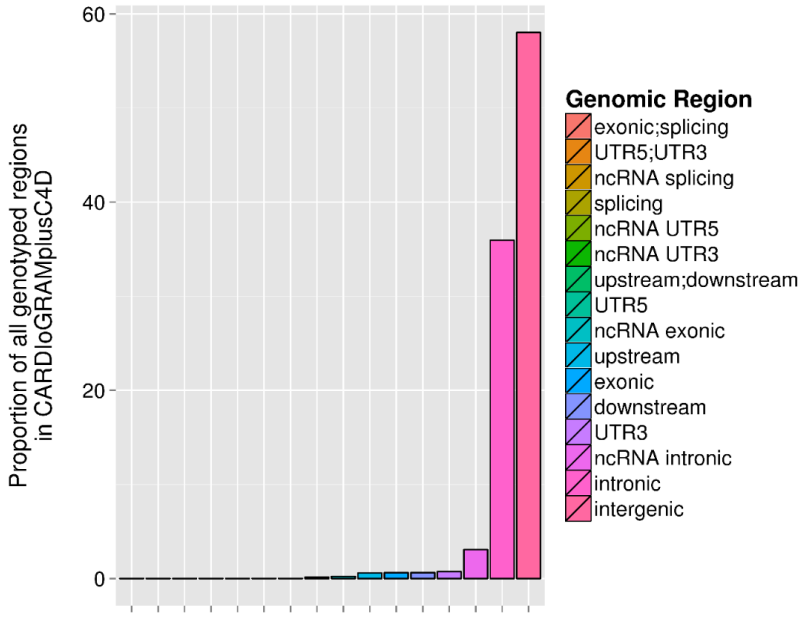


Figure 1.2: Breakdown of genomic region annotations for 9.4 million variants tested in the CARDIoGRAMplusC4D meta-analysis. Abbreviations are as follows: UTR5 = 5-prime untranslated region; UTR3 = 3-prime untranslated region; ncRNA = non-coding RNA.

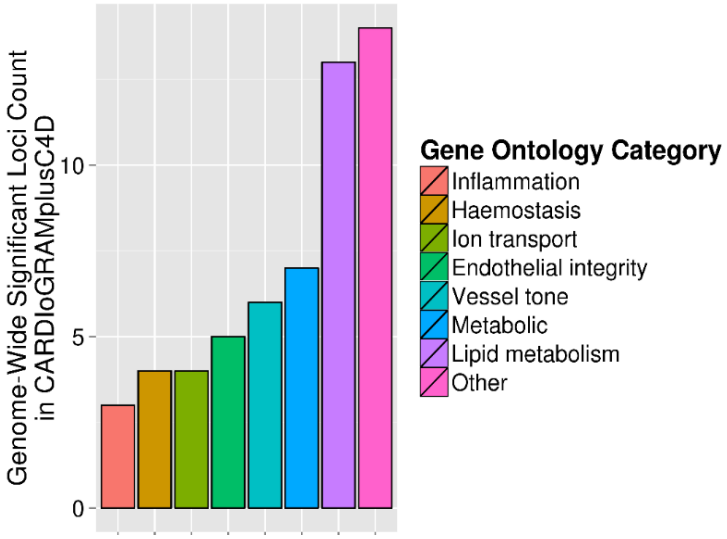


Figure 1.3 Functional distribution of 58 genome-wide significant loci from the CARDIoGRAMplusC4D consortium.

1.4 Exome sequencing and rare variants

GWAS have proved invaluable in enhancing our knowledge of the interactions between common variants and complex disease through the use of microarray-based genotyping. However, traditional GWAS have made little contribution to understanding the population genetics features of rare variation and the role of rare variants in complex disease development. Since it has been previously shown that rare variants collectively outnumber common variants and are enriched within the protein-coding regions (exons) of the genome, exome sequencing has emerged as a powerful tool for rare variant discovery and the standard laboratory for rare variant association studies [14–16].

The development of exome sequencing strategies allows for both known and novel rare variations to be discovered within the highly conserved regions of the genome by virtue of *de novo* variant calling. Exome sequencing selectively captures the DNA sequence corresponding to coding regions, which make up approximately 1.5% of the genome's 3 billion bases [17,18]. Despite this low proportion, the exons contain approximately 85% [17] of all disease-causing mutations which make them ideal sources to find causal variants that directly contribute to manifestation of complex diseases like CAD. Moreover, it has been shown that rare variants are enriched for protein-altering mutation types such as nonsynonymous variation (single nucleotide variants that result in an amino acid substitution within the protein sequences) and frameshift indels (insertions/deletions that alter the reading frame of the transcript) which hold great clinical significance as they are more likely to alter or perturb protein function, allowing them to exhibit higher penetrance within case populations [14-16].

Therefore, it is proposed that rare variants hold significant biological relevance and can contribute to furthering our understanding of the molecular pathways and gene networks involved in the manifestation of complex disorders such as CAD.

1.5 Defining and implementing rare variants in association analysis

The rationale for assessing the contribution of rare variants on CAD has primarily emerged from three arguments: 1) the exponential population boom over the past millennium has led to an abundance of novel, rare variation compared to common variants across the genome, 2) the vast majority of damaging variants are rare due to the influence of purifying selection and 3) common variants have not accounted for total disease heritability, despite the efforts of very large and well-designed case-control meta-analyses. Taken together, these factors provide sufficient grounds to evaluate the effect of rare variants on CAD risk in order to determine whether they can account for missing heritably. However, in order to effectively employ rare variants in epidemiological analyses, it is essential to 1) accurately discriminate genuinely rare variants from those that are cryptically rare by virtue of small samples sizes and 2) employ appropriate statistical tests that are well powered to assess the putative associations between rare variants and the phenotype of interest.

The recent developments of the Genome Aggregation Database (gnomAD) [19] and Exome Aggregation Consortium (ExAC) [20,21] have allowed investigators to accurately determine which variants are truly rare in their study cohorts by leveraging the immense sample size of these publically available consortia datasets (N = 60706 for ExAC and N = 123236 for gnomAD). Specifically, by identifying the corresponding allele frequencies of

a given variant in gnomAD/ExAC, one is able to robustly demarcate genuinely rare variants and filter out variants that are actually common once inspected in large sequencing databases. Moreover, both gnomAD and ExAC have stratified allele frequencies across five major global ethnicities (Non-Finnish Europeans, Africans, South Asians, East Asians, and Latin Americans) and two founder populations (Finnish Europeans and Ashkenazi Jews) which will prevent spurious rare variant associations due to population stratification. Additionally, sequence consortia datasets allow investigators to experiment with differing allele frequency thresholds to define a "rare" variant in their cohort. These thresholds are expected to vary depending upon the mutation types of the variants under investigation (i.e. missense, nonsense, frameshift etc…) and the mode of inheritance model (i.e. additive, dominant, recessive). However, most typically, rare variant are defined having a MAF of less than 1-5% in the general population.

Due to rare variants being individually very infrequent, they cannot be assessed using single locus association methods that are typically used for associating common variation with disease phenotype in GWAS. Specifically, single locus association tests such as chi-squared (χ^2) conducted on individual rare variants in a case control analysis would result in a significant deflation of test-statistic resulting in p-values that fail to reach even nominal significance (i.e. $P < 0.01$) [22]. Therefore, single locus association analyses are vastly under powered in the context of individual rare variants.

One method by which rare variant association analysis can be conducted with high power (proposed by Li and Leal, 2008) [22] is if they are analyzed *in aggregate* across a specified genetic context (i.e. an individual gene or a set of genes involved in a single

pathway) (Figure 1.4). The *in aggregate* method represents a rare variant burden test that accounts for the inherent infrequency of individual rare variation by summing the allele counts for a specified set of rare variants (e.g. within a gene) across all samples under investigation (equation 1.1; where B_i represents the cumulative count of rare variants per individual given their genotype G_i across j variants in a gene). This approach allows rare variants to undergo association with disease phenotype collectively rather than independently which may allow for the discovery of genes with high evolutionary constraint.

$$B_i = \sum_{j=1}^{M} G_{ij} \qquad \text{Equation 1.1}$$

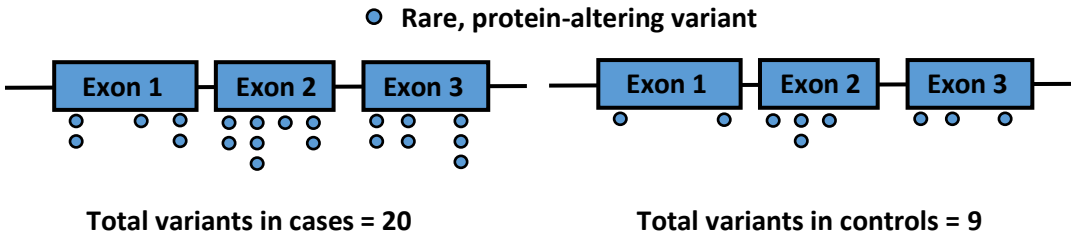


Figure 1.4: *In aggregate* method of rare variant burden testing. Individual MAF for rare variants meeting specified "deleterious" criteria are summed to a single test statistic in both cases and controls and subsequently associated with disease phenotype.

Figure 1.4 provides an arbitrary example of the *in aggregate* burden method where rare variants meeting a pre-specified "deleterious" criteria (in this case, any variant that alters the structure of the protein) are combined across a single gene (containing 3 exons) in cases and controls. Here, 20 rare, deleterious variants are observed in cases compared to only 9 in controls, resulting in a large effect size (OR 2.2).

While burden tests provide a novel approach for assessing the collective effect of rare variants on disease risk, they hold the assumption that all variants in a given region or pathway exert their effect in a single direction. That is, variants that exhibit a neutral or protective effect on a given phenotype are regarded as causal risk alleles in the association model. This limitation has spurred the development of variance component tests which account for direction of effect by regressing *individual* rare variants on the phenotype of interest instead of combing them into an aggregate test statistic to be used in association testing [23]. The distribution of test statistics obtained from multiple regressions of individual rare variants can be compared to a null distribution to determine statistical significance. This approach avoids the *a prori* assumption of directionality and significantly increases power of rare variant association when it is expected that target regions harbour a mixture of protective, neutral, and risk alleles.

1.6 Notable examples of rare variant successes

The distinct nature of rare variants can lend itself towards the discovery of novel genes that mediate disease risk outside of the common variant common disease paradigm and can result in the elucidation of novel biological mechanisms underlying disease development. This is effectively illustrated by a large, case-control study of early MI conducted by Do *et al.* 2015 [2]. In this analysis, whole-exome sequencing was used to identify the association between the burden of rare, protein-altering variants within Apolipoprotein A-V (APOA-V) and onset of early CAD/MI ($p = 5x10^{-7}$; OR = 2.2) [2]. Previous to this analyses it was well established that common variants within the *APOA-1-APOC-III-APOA-IV-APOA-V* locus had been associated with CAD as observed in the

CARDIoGRAMplusC4D consortia (OR = 1.16) [5,12]. However, due to the extensive linkage disequilibrium (LD) at this locus, the causative gene could not be defined. Since rare variants are inherently infrequent, they are unlikely to exist in LD amongst each other or with common variants. Therefore, the burden of rare variants within a particular gene in this locus provides strong rationale for it being the main mediator of disease risk. Furthermore, carriers of rare, protein altering variants in APOA-V were also found to exhibit increased triglyceride levels, which underscores the importance of accounting for triglycerides as a strong, modifiable risk factor for CAD/MI and weakens the notion that development of atherosclerotic plaques are driven solely by LDL cholesterol. Since publication of this work, a candidate-gene exome sequencing analysis conducted by Khera *et al.* 2017 further demonstrated that the burden of rare, protein-altering mutations within Lipoprotein lipase (LPL) were also associated with both increased triglycerides and early onset CAD/MI [24].

Additionally, rare variant association analyses recently resulted in the discovery of a novel gene associated with LDL cholesterol. Specifically, investigators conducted whole-exome sequencing in phenotypic extreme cohort comprised of individuals within the 99th and 1st percentile of LDL cholesterol as cases and controls, respectively [25]. A total 8 gene-based burden and variance component tests were conducted for rare, protein-altering variants and revealed a significant association between Patatin-like phospholipase domain-containing 5 (PNPLA5) with LDL cholesterol at high effect ($p = 3x10^{-7}$; Beta = 1.2 mmol/L) [25]. The PNLPA5 locus was not found to reach genome wide significance in the Global Lipids Genetic Consortium (GLGC) GWAS and probably modifies LDL

cholesterol levels through the presence rarer mutations which is likely a consequence of higher evolutionary constraint for this gene. Interestingly, the burden of rare variants in PNPLA5 along with 3 other previously known lipid genes determined to be significant (LDLR, APOB, PCSK9) resulted in a combined LDL heritability estimate of 5.4% which is quite extensive when considering that common variants across 18 novel loci associated with LDL in the GLGC only accounted for 2.6% of LDL heritability [25].

Taken together, the study results indicate that rare variant association analyses can lead to the discovery of novel disease etiology and identification of new risk genes which can refine diagnostic criteria for CAD and improve individual risk stratification for prognostic measures, respectively.

1.7 Identifying protective effects of damaging mutations: A noteworthy application of whole-exome sequencing

Exome sequencing has mediated the discovery of numerous genes that confer risk for CVDs through loss-of-function (LOF) mechanisms [2,26,27]. However, certain genes harbouring LOF mutations confer protection against complex diseases which has accelerated the development and clinical implementation of therapeutic inhibitors to prevent primary disease outcomes for high risk individuals. Most recently, numerous clinical trials have been shown to robustly reduce incidence of CAD and dyslipidemia in a manner that directionally mimics LOF mutations in the targeted gene, thus establishing a sound correlate between pharmacological intervention and genetic predisposition [28]. However, the impact of lifelong exposure to LOF mutations on CAD risk can only be identified through genetic epidemiology. To date, there have been three genes identified

through exome sequencing in which LOF mutations confer protection against CAD: (1) Niemann–Pick C1-like 1 (*NPC1L1*) [29], Apolipoprotein-C3 (*APOC3*) [30], and Angiopoietin like-4 (*ANGPTL4*) [31]. Large, multi-ethnic case-control study designs were used to identify significant enrichment of LOF alleles within these genes amongst individuals with no presentation of dyslipidemia or CAD. For the purposes of this work, only APOC3 will be discussed as this discovery was a result of whole-exome sequencing with no *a prioi* hypothesis. However, a summary of all three genes is provided in Table 1.3.

1.7.1 APOC3

APOC3 is a peripheral membrane protein located within the *APOA-1-APOC-III-APOA-IV-APOA-V* locus, which has robustly been shown to confer risk for CAD in large genome-wide meta-analyses [5]. APOC3 synthesis occurs largely within hepatocytes and is shown to associate with both triglyceride-rich lipoproteins (TRLs) (e.g. chylomicrons, very-low density lipoprotein) and HDL. APOC3 is a robust regulator of triglyceride homeostasis by acting as an inhibitor of LPL, an enzyme that catalyzes the hydrolysis of triglycerides to free fatty acids along the capillary endothelium [30,32]. Additionally, APOC3 acts in an LPL-independent manner to prevent hepatic uptake of TRL remnants by disrupting their association with remnant receptors [30,32].

Due to the evidenced heritability of plasma triglycerides and their correlation with CAD, the TG and HDL Working Group of the National Heart, Lung, and Blood Institute Grand Opportunity Exome Sequencnig Project v6500 (NHLBI-GO ESP6500) sought to further demarcate the genetic architecture of triglyceride homeostasis by sequencing the protein-coding regions genome in 3734 individuals of European or African ancestry across

7 cohorts with relevant phenotypic data for cardiovascular disease. The investigators restricted analysis to rare (MAF < 1%) SNVs resulting in missense, nonsense, or splice site mutations [30]. In the discovery, investigators determined the gene-based burden of rare mutations within APOC3 was most strongly associated with plasma triglycerides in Europeans ($P=7x10^{-6}$) and Africans ($P=1x10^{-5}$) after adjustment with principle components of ancestry [30]. Heterozygote carrier frequency of one or more rare mutations found in discovery sequencing was 1/150 where carriers had a 39% reduction in mean plasma triglycerides ($P=6x10^{-9}$) compared to non-carriers [30]. The rare variants contributing to the gene-based association signal in discovery whole-exome sequencing (one missense, one nonsense and two splice-site) were further genotyped in 34,002 CAD patients and 76,968 disease-free controls of European, African, and Hispanic ancestry across 15 studies within the NHLBI-GO ESP 6500. A total of 498 individuals (113 cases and 395 controls) were found to be heterozygous for at least one of the genotyped variants and a 40% decreased risk for CAD compared to non-carriers (OR 0.60; $P=4x10^{-6}$) after adjusting for principle components [30].

An antisense inhibitor of APOC3 mRNA (ISIS 304801) has recently demonstrated promising clinical efficacy in treating patients suffering from severe hypertriglyceridemia. Gaudet *et al.* 2014 recruited three patients presenting with triglyceride levels ranging from 15.9 to 23.5 mmol/L [30]. All patients harboured either a homozygous or trans compound heterozygous mutation in LPL which compromised the catalytic activity of each copy by 95% [30]. Patients were administered ISIS 304801 at a 300 mg dosage once a week over a 13 week period and were evaluated in comparison to the initial baseline measurements. At

end of treatment, plasma triglyceride levels declined by 56 to 86% across all three patients with 2/3 patients reaching levels as low as 2.6-2.8 mmol/L during the treatment period [30]. These findings, when taken together with the aforementioned association between APOC3 and CAD, provide tremendous promise that pharmacological modification of plasma triglycerides may be preventative for CAD development or associated complications.

Table 1.3: Summary descriptions of genes conferring protection against CAD when harbouring LOF mutations. P-values and odds ratios correspond to the association between mutation carrier status and CAD.

Gene	Mutation carrier frequency in CAD cases/controls (%)	P-value	Odds ratio (95% CI)	Pharmacological inhibitor	Drug type	Effect of pharmacological inhibitor in clinical trials ‡
NPC1L1	0.04/0.09	0.008	0.47 (0.25-0.87)	Ezetimibe	Small molecule inhibitor	6.4% risk reduction for CVD death, MI, unstable angina (with hospitalization), coronary revascularization, or stroke.
APOC3	0.3/0.5	$4x10^{-6}$	0.60 (0.47-0.75)	ISIS 304801	Small inhibitory RNA	56 to 86% decrease in plasma triglycerides.
ANGPTL4	0.07/0.13	0.04	0.47 (NA)	REGN1001	Monoclonal antibody	42% and 31% decrease in plasma triglycerides for mice and non-human primates, respectively.

‡ percent values represent difference in relative risk between control and treatment groups

1.8 Pathophysiology of CAD

The molecular mechanisms underlying CAD pathogenesis can offer tremendous value in facilitating the prioritization of statistically significant genes associated with a disease phenotype. This will ultimately serve to discriminate genes that have a potential role in disease progression from those that are statistical artefacts.

The development of CAD can largely be attributed to disturbances in homeostatic maintenance of inflammatory and metabolic pathways that collectively contribute to the formation of atherosclerotic plaques within the coronary arteries. Plaque development is a highly progressive process with clinical complications such as angina and myocardial infarction typically arising in men > 55 and women > 65 years of age [33]. Plaque formation begins when the arterial endothelium encounters excessive concentrations of plasma Low-Density Lipoprotein Cholesterol (LDL-C), which is responsible for transporting endogenous cholesterol from the liver to extra-hepatic tissues [34]. Upon reaching a variable threshold concentration within the vasculature, plasma LDL-C will transcytose through the arterial endothelial monolayer and enter the sub-endothelial space known as the *tunica intima*, which represents the foci of atherosclerotic plaque development [33,34]. Following transcytosis, LDL-C particles become vulnerable to enzymatic, cell-mediated oxidation, primarily from myeloperoxidases and lipoxygenases expressed by resident macrophages and endothelial cells, respectively [33–35]. Oxidized LDL-C (oxLDL-C) is highly atherogenic as it contributes to chemo-attraction of pro-inflammatory cells, endothelial cell dysfunction, and vascular smooth muscle cell (VSMC) proliferation which are all necessary processes for establishing the initial stages of plaque development.

1.8.1 Chemo-attraction of pro-inflammatory cells

Resident macrophages within the *tunica initma* express scavenger receptors such as scavenger receptor class A member I/II (SCARAI/II) and scavenger receptor class B member I (SCARB1) that facilitate phagocytosis of cellular debris, apoptotic cells, and foreign organisms [33]. These receptors have shown to display high affinity to oxLDL-C compared to standard LDL-C due to the increased hydrophilic incurred upon cell-mediated enzymatic oxidation. The scavenger receptors facilitate the uptake of oxLDL-C which causes a significant increase in their intracellular lipid content and results in intimal macrophages adopting a "foam"-like appearance (foam cells) [33,34]. This phenotypic change is accompanied by increased secretion of pro-inflammatory cytokines such as macrophage-colony stimulating factor (mCSF) and monocyte chemo-attractant protein 1 (MCP-1) which act collectively to recruit pro-inflammatory monocytes to the early plaque foci [33]. Upon reaching the intimal layer, recruited monocytes readily differentiate into macrophages and further internalize oxLDL-C to become foam cells which accumulate and contribute to the formation of a fatty-streak.

1.8.2 Endothelial cell dysfunction

The pro-inflammatory milleau generated by migration of innate immune cells into the arterial intima induces the expression of lectin-like oxidized receptor 1 (LOX1) on the surface of endothelial cells [36]. LOX1 is able to preferentially bind oxLDL-C ligand which results in a nuclear factor kappa-light-chain-enhancer of activated B cells (NF-kB) mediated intracellular cascade, leading to transcriptional activation of endothelial adhesion molecules such as intercellular adhesion molecule 1 (ICAM-1) and vascular-cell adhesion

molecule 1 (VCAM-1) [37]. These adhesion molecules mediate the process of extravasation of pro-inflammatory monocytes which involves rolling adhesion of these cells across the surface endothelial layer followed by paracellular transport into the arterial intima.

Additionally, the NF-kB mediated intracellular cascade, induced by LOX1-oxLDL-C interaction, enhances endothelial transcription of chemokine c-c motif ligand 3 (CCL3) and chemokine c-c motif ligand 4 (CCL4) chemoattractants which are able to recruit type 1 T helper (TH1) cells to the arterial intimal via extravasation [36]. TH1 cells play an active role in exacerbating the atherosclerotic plaque and are largely responsible for the manifestation of clinical complications such as MI (discussed in section 1.7).

1.8.3 VSMC proliferation and migration

VSCMCs exhibit both contractile and synthetic properties as they are actively involved in maintaining vasomotor tone and synthesizing extracellular matrix (ECM) proteins such as elastin and collagen which provide the arteries with structural integrity. Presence of oxLDL-C has been shown to induce VSMC proliferation and migration from the *tunica media* to the intimal layer where they participate in the formation of a fibrous cap that surrounds the lipid-rich fatty streak. Specifically, oxLDL-C stimulate secretion of growth factors such platelet derived growth factor (PDGF) and basic fibroblast growth factor (bFGF) from endothelial cells which collectively shift VSMCs from a quiescent, contractile state to a proliferative, migratory phenotype whilst up-regulating their capacity to synthesize ECM proteins (mainly Type IV and V collagens). The increased deposition of collagen within the arterial intima results in the formation of a fibrous cap that envelopes the fatty streak and prevents it from advancing into the arterial lumen. However, as the

fibrous cap progressively enlarges, it can begin to narrow the luminal area and occlude blood flow, leading to localized tissue ischemia and subsequent infarction if the plaque is sufficiently advanced.

1.9 Mechanisms of Acute Coronary Syndromes

Acute Coronary Syndromes (ACS) refer to any set of clinical outcomes consistent with coronary vessel occlusion and myocardial ischemia [34]. There are three primary outcomes defined by the American College of Cardiology (ACC) that demonstrate high prevalence amongst individuals with CAD: unstable angina (UA), non-ST-elevated myocardial infarction (NSTEMI) and ST-elevated myocardial infarction (STEMI) [38]. These outcomes range in clinical severity in terms treatment strategies and patient prognosis (Figure 1.5A) and typically manifest based on both the degree of coronary vessel occlusion and structural integrity of the fibrous cap.

UA and NTEMI are collectively categorized as non-ST-elevated acute coronary syndromes (NSTEACS) and result from transient or intermittent obstruction of blood flow due to narrowing of the arterial lumen by progressive plaque growth [38]. NSTEACS manifest as irregular pains that occur in the chest, arm, and/or jaw in the absence of physical exertion [35]. Additionally, the pain does not typically alleviate with rest or short-term pharmacological interventions such as calcium-channel blockers (diltiazem) or vasodilators (nitroglycerin) [35,38]. The most prominent distinction between UA and NSTEMI is the presence of plasma cardiac troponin I (cTnI) and T (cTnT), which are diagnostic of more extensive myocardial ischemia [35].

Unlike NTSEACS, cases of STEMI are attributed to sudden death or immediate need of clinical involvement through percutaneous coronary intervention or coronary bypass surgery [35]. This is largely a consequence of unstable fibrous caps which can rupture and expose lipid-laden core of the atherosclerotic plaque to the clotting factors circulating in the arterial lumen [35,39]. Due to the thrombogenecity of the lipid core, contact with clotting factors will lead to the formation of a thrombus at the rupture site which will frequently cause complete vessel occlusion and manifest as a clinically severe acute ischemic event (Figure 1.5B) [40]. Atherosclerotic plaques with high macrophage and TH1 content tend to be particularly susceptible to plaque rupture as cross-talk between these cells mediate the degradation of interstitial collagens via matrix-mettaloproteinases (MMPs) [33,39]. MMPs are proteolytic enzymes that exhibit high affinity toward interstitial collagens and are overexpressed in inflammatory states due to the binding of CD40 ligand (TH1 derived) to CD40 expressed on the cell surface of macrophages [40]. This interaction can induce increased macrophage expression of MMPs which ultimately results in the degradation of the fibrous cap, rendering it vulnerable to rupture.

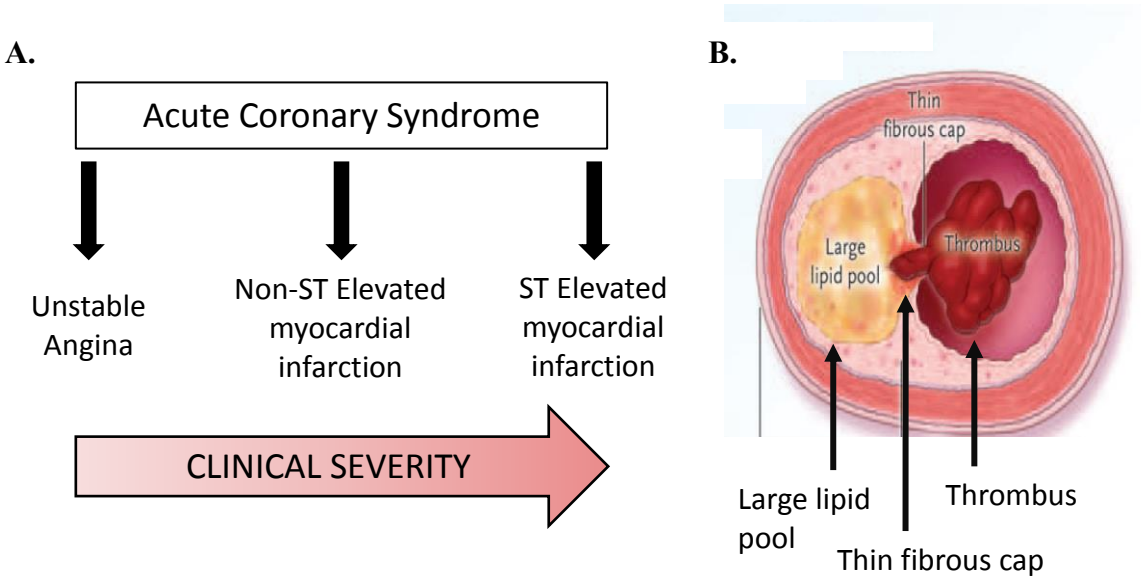


Figure 1.5: Summary of ACS. (A) Schematic of the continuum of clinical severity for types of ACS observed as a result of CAD. (B) depiction of thrombus formation due to a thin fibrous cap resulting in complete vessel occlusion and STEMI (B). Panel B was taken from Libby *et al.* 2013 [40].

CHAPTER 2 – DECODE: Project Proposal for an Early-Onset Coronary Artery Disease Cohort

2.1 Introduction

Seminal epidemiological analyses have consistently demonstrated that genetic factors profoundly impact the risk for developing CAD among young individuals [8–10]. These observations are founded on the principle that patients presenting with CAD at an early age are unlikely to have been chronically exposed to common risk factors that may have otherwise mediated disease onset (described further in section 1.1). As such, considerable interest has been drawn towards conducting extreme phenotype sampling (EPS) where individuals are selectively recruited from the tails of a given continuous phenotypic distribution (e.g. age) in order to enrich for rare causal alleles in CAD cases and protective alleles in disease-free controls. Therefore, EPS can lead to substantial increases in statistical power to detect rare variant association signals [41,42] which would otherwise have to be addressed through exorbitantly large sample sizes.

Power estimates are also expected to increase through the biological validation of putative disease-causing variants within candidate genes that participate in molecular pathways known to be involved in CAD pathogenesis [43]. Specifically, the effects variants discovered through next-generation sequencing technologies can be ascertained by functionally profiling their effects in *in vitro* cellular models as opposed to relying on *in silico* pathogenicity scores which cannot always accurately discriminate between neutral and risk alleles, especially among missense variants. Therefore this strategy allows for the stringent filtering of non-functional neutral alleles that dampen rare variant association signals by leveraging basic science techniques of site-directed mutagenesis and overexpression

Therefore, employing both EPS and biological validation techniques in rare variant association studies of CAD will vastly empower the ability to identify rare causal variants and establish sound phonotypic correlates at the molecular level.

We herein propose our pilot study – DECODE, a comprehensive investigation of an early-onset CAD (EOCAD) cohort that will undergo whole-exome sequencing to identify rare variants of high effect that can be biologically profiled in order to identify genes that are causally linked to CAD. Through this endeavor, we aim to refine our understanding of the genetic determinants underlying CAD and account for the heritability not currently explained by common variants.

2.2 Hypothesis

1.) We hypothesize that rare genetic variation within the protein coding regions of the genome play a vital role in CAD development, especially in individuals with very early CAD. We propose that genomic and in vitro cellular analyses of early-onset CAD-case samples can discover genes harbouring these variations and elucidate their biological consequences by leveraging extreme phenotypic sampling and the use of well-defined external and internal control datasets.

2.3 Primary objectives

1.) To identify and biologically characterize rare, protein-altering genetic mutations responsible for very early CAD using burden and variance component testing under case-only and case-control study designs. DECODE study participants will be consented and their blood will be drawn for downstream genomic, macrophage and

stem cell analysis. These analyses will act to determine the genetic contributors to early onset CAD and assess their phenotypic and functional consequences in the context of vascular abnormalities.

2.) To develop per-sample correction factors that can be used to calibrate 'N of 1' benchmarking analyses in practical applications such as rare variant association testing and calculation of rare variant gene scores.

3.) To determine the prevalence of monogenic dyslipidemias (especially familial hypercholesterolemia) in young, angiographically-proven CAD patients by evaluating rare, protein-altering variants in known genes. In doing so, we aim to establish criteria for familial hypercholesterolemia screening in early CAD patients.

2.4 Methods

2.4.1 Study population

A total of 55 participants were recruited into the DECODE study through both Hamilton Health Sciences Heart Investigation Unit and Lipid Clinic from September 2014 to May 2016. In order qualify for inclusion, all individuals had to exhibit angiographically-proven CAD with at least 70% stenosis in a single coronary vessel at age 40 or under for males and 45 or under for females. Individuals with co-morbidities including chronic kidney disease (CKD), type I diabetes mellitus, insulin-dependent type II diabetes mellitus, chronic hepatitis, HIV, vasculitis, systemic autoimmune disease, or chronic consumers of amphetamine/steroids were deemed has having a secondary cause of CAD and were not further considered as study candidates. This study has received annual approval from the

Hamilton Research Ethics Board since conception and is in full compliance with the Declaration of Helsinki. All eligible individuals provided written consent to participate in this study.

2.4.2 Blood collection protocol

All DECODE study participants underwent cardiac catheterization at the Hamilton Health Sciences Heart Investigation Unit. 10mL whole-blood was drawn into 3 purple-top EDTA vacutainers for genomic, stem cell and macrophage analysis. Blood was also drawn into PaxGene RNA tubes for gene expression analysis (Figure 2.1). For the genomic arm, DNA was extracted from peripheral blood leukocytes using the QIAsymphony SP with the QIAsymphony DNA mini kit **(**QIAGEN**)**. Samples were thereafter quantified on the Qubit 2.0 fluorimeter using Qubit dsDNA High Sensitivity assay kit **(**Life Technologies**)** prior to library preparation for sequencing.

2.4.3 Ion Ampliseq™ exome library preparation

PCR-based exome enrichment was used to prepare exome libraries using the Ion AmpliSeq™ Exome RDY Panel and Library Kit Plus with HiQ chemistry (Life Technologies). Briefly, unique 2X Exome Primer Pools and PCR master mix (5X Ion AmpliSeq™ HiFI Mix, Nuclease-free Water) were added to 100 ng of starting DNA across 12 wells and amplified on the Veriti 96-well Thermal Cycler (Life Technologies) using standard parameters defined for Ion Ampliseq™ enrichment (Table 2.1). PCR reactions were pooled into single wells to generate per sample whole-exome libraries. Library amplicons were ligated with Ion Xpress™ barcode and P1™ adapter sequences for cross-

sample pooling and hybridization to Ion Sphere Particles (ISPs), respectively for template preparation. Ligated exome libraries were normalized to 100 pm after quantification using quantitative PCR (qPCR) on the ViiA 7 platform (Life Technologies) using the Ion Library TaqMan® Quantitation Kit.

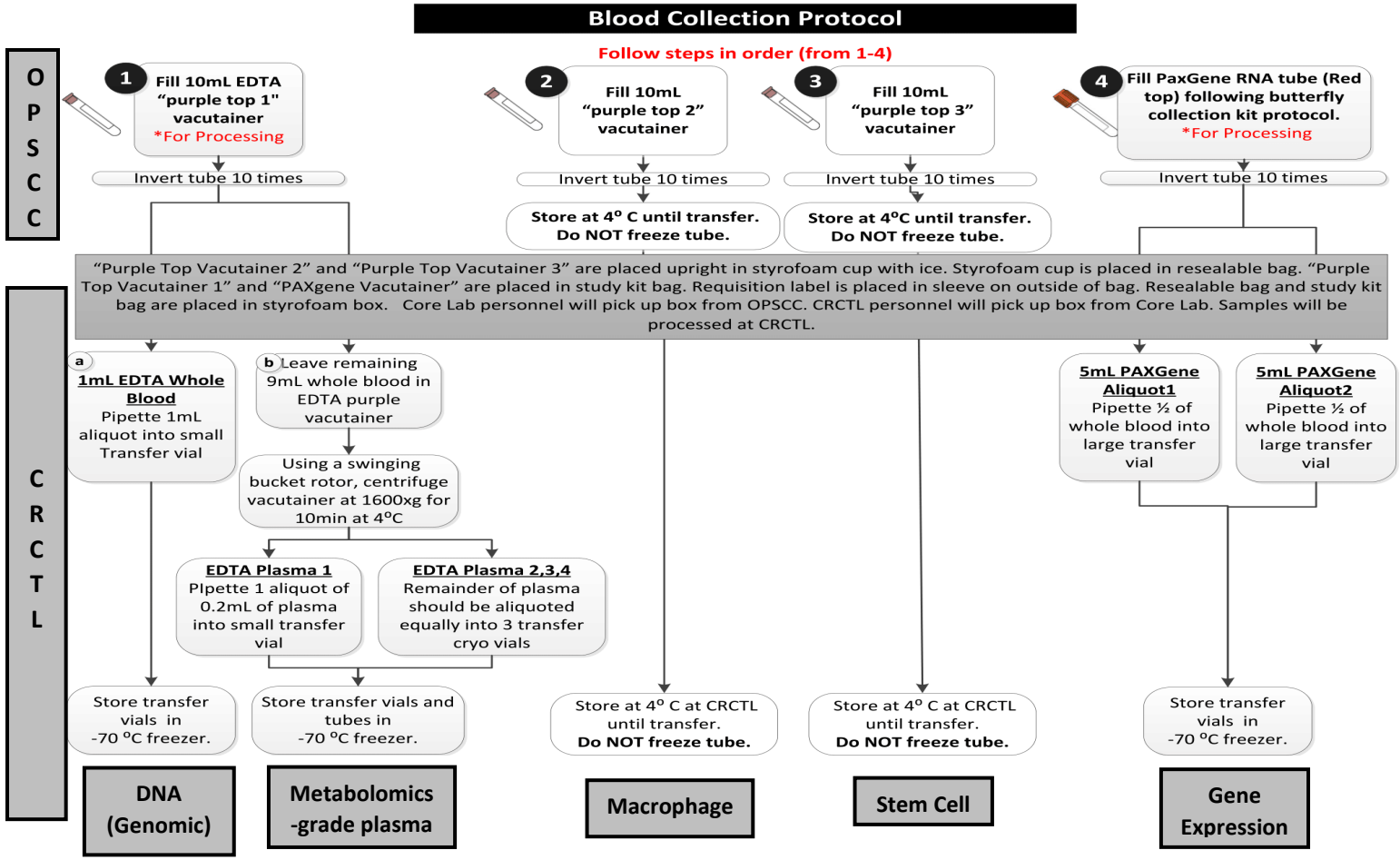


Figure 2.1: Blood collection workflow for the genomic, stem-cell, macrophage, and gene-expression arms of the DECODE study. Blood collection and specimen processing/storage was performed by OPSCC and Clinical Research and Clinical Trial (CRCTL) personnel, respectively at the Hamilton General Hospital.

Table 2.1: PCR run parameters for Ion Ampliseq™ exome enrichment.

Stage	Temperature (°C)	Time (minutes)	Cycles
Polymerase activation	99	2:00	1
Denaturation	99	0:15	
Elongation	60	16:00	10
Annealing	60	16:00	

2.4.4 Template preparation

For sequencing on the Ion S5XL™ instrument, barcoded exome libraries were pooled in groups of 2 and loaded onto the Ion Chef ™ platform with Ion 540TM Chef Reagents and Ion S5TM Solutions (Life Technologies). Here, library amplicons hybridize with ISPs and are clonally amplified with emulsion PCR using biotin-complexed primers. Template-positive ISPs are enriched using streptavidin bead pull down and are treated with NaOH (8M) to remove the complimentary strand bound to the biotin-complexed primers. Single-stranded DNA template-ISP complexes were thereafter loaded in 2 plex onto the Ion 540™ chip (Life Technologies). Samples sequencing on the Ion Proton™ instrument underwent template preparation using the Ion PI™ Template OT2 200 Kit v3 and the Ion One Touch 2 system (Life Technologies). Single-stranded DNA template-ISP complexes were loaded in 2-3 plex onto the Ion P1™ chip (Life Technologies).

2.4.5 Exome sequencing and read mapping

Prepared template DNA underwent semiconductor-based long-read sequencing on the Ion Proton™ and Ion S5XL™ platforms (Life Technologies) at the Genetics and Molecular

Epidemiology Laboratory at the David Braley Cardiac, Vascular and Stroke Research institute. Prior to variant calling, reads are pre-processed to trim adapter sequences and low-quality base calls at their 3' ends since higher quality base calls tend to be concentrated closer to the 5' end where the flow signal is strongest. Quality trimming is conducted using phred-scaled, per-base quality scores and initiates once these quality scores fall below a pre-specified threshold. Bases falling below this threshold will thereafter be excluded from downstream analysis. Trimmed reads that exhibit a short length (< 8 bp), contain adapter dimers, or lack sequencing keys are filtered from analysis in order to facilitate their downstream alignment. Additionally, polyclonal reads (i.e. reads from multiple templates on a single ISP) are filtered out in order to optimize flow signal strength.

Pre-processed reads (mean ~250bp) were aligned to the GRCh37/hg19 human reference genome assembly using version 5.2 of the torrent mapping aligning program (TMAP 5.2) (Life Technologies). TMAP functions by first generating a composite list of Candidate Mapping Locations (CML) using various subsets of four established alignment algorithms (BWA-short, BWA-long, Sequence Search and Alignment Hashing Algorithm, and Super-maximal Exact Matching). The CML lists are used to align reads to the reference assembly using the Smith Waterman Algorithm in order to obtain multiple alignment sets which can be aggregated together to elucidate the alignment set with the highest overall mapping quality. The most optimal alignment set represents the Binary Alignment (BAM) file that is used as the input to evaluate candidacy of potential variants.

Following read alignment, duplicate reads are typically marked and removed using and picard *mark duplicates* and samtools *remove duplicates*, respectively. However, reads

generated using Ion Ampliseq™ Exome panel should not be undergo duplicate marking and removal as the reads are expected to have the same start sites due to PCR-based exome capture. As such, duplicate removal could potentially result in an underestimation of coverage at a given variant site, which could ultimately compromise the overall sensitivity of the final variant call set.

2.4.6 Sequencing quality control

Per-base depth of coverage values were computed using the GATK DepthOfCoverage tool across Ion Ampliseq™ Exome target regions. In house shell scripts were subsequently used to calculate 3 coverage-based metrics to assess the quality of a given exome: 1) mean depth of coverage across target bases, 2) proportion of target bases covered by at least 20 reads (% > 20X coverage), and 3) the proportion of target bases covered by at least 0.2% of the mean depth of coverage (coverage uniformity). Samples that were lower bound outliers (< Q1-1.5xIQR) in any of these metrics after initial sequencing were re-sequenced and combined with their previous exome in order to achieve higher quality. Additionally, any samples achieving < 75% 20X coverage were also re-sequenced and combined with their previous exome regardless if they were outliers in this category or of their performance in the other coverage-based sequencing metrics. 2 additional metrics were used to gauge sequencing quality independent of coverage: 1) proportion of reads mapped to target regions (% on-target reads), and 2) proportion of bases with a phred scaled quality of at least 20 (% Q20 bases). Lower-bound outliers in these 2 categories underwent both library preparation and sequencing again and kept independent of their previous exome. Both non-

coverage-based metrics were obtained directly from the ion torrent browser. All metrics were stratified according to sequencing platform (Ion Proton™ or Ion S5XL™), number of samples loaded onto sequencing chip (3plex or 2plex), and template preparation strategy (IonChef™ or IonOneTouch™).

2.4.7 Variant calling

Single nucleotide variants (SNV) and insertion/deletions (INDEL) were called with version 5.2 of the Torrent Variant Caller (TVC 5.2) (Life Technologies) from pre-processed BAM files according to variant filtering parameters defined in germline, low-stringency settings by Ion Torrent (Table 2.2). Briefly, genotype likelihoods were assigned for candidate variants and were used to compute the posterior probability of a variant genotype (heterozygous (0/1) or homozygous alternate (1/1)) using Bayesian or Frequentist methods. Variants that were successfully genotyped were subsequently filtered if they exhibited low depth of coverage (DP ≤ 5.0 SNV; DP ≤ 10.0 INDEL), low phred-scaled quality (QUAL ≤ 15.0 SNV; QUAL ≤ 20.0 INDEL), low alternate allele count to read ratio (AF < 0.1 SNV; ≤ 0.25 INDEL), high strand bias (STB ≥ 0.98 SNV; STB ≥ 0.90 INDEL), mapped to large homopolymer regions (HRUN ≥ 8 SNV & INDEL), high degree of signal shift (RBI ≥ 0.25 SNV & INDEL), or low relative rad quality (MLLD ≤ 5 SNV & INDEL).

Homozygous reference calls were generated by creating a hotspot variant call file (VCF) from the previously combined callset using the TVC utils prepare_hotspot function. Variants were individually re-called across all samples using the hotspot VCF as an input

file. Variants present within the hotspot VCF but absent in the individual sample VCF were assigned either a missing (./.) or homozygous reference (0/0) genotype.

Variant data for each sample was formatted in variant call format (VCF) to display the chromosome, genomic position (hg19), reference allele, alternate allele, variant phred-scaled quality, pass/fail status, sequencing metrics and genotyping metrics for all variants. Variants were identified and genotyped individually across all samples and then merged into a single VCF file using the Genome Analysis Toolkit's (GATK) [44] CombineVariants tool.

Table 2.2: Quality metrics and thresholds used in germline, low stringency settings to filter out low-quality variant calls

TVC filtering parameters	SNV threshold	INDEL threshold	Short description
Depth of coverage (DP)	≤ 5	≤ 10	Number of reads aligned to candidate variant site
Phred-scaled variant quality (QUAL)	≤ 15	≤ 20	The probability that a candidate variant is not an error
Alternate allele count to read ratio (AF)	≤ 0.10	≤ 0.25	Ratio of number of reads calling a variant allele to the total reads aligning to candidate variant site
Strand bias (STB)	≥ 0.98	≥ 0.90	Reads with variant allele are mapped disproportionately to either forward or reverse strand
Homopolymer run (HRUN)	≥ 8	≥8	Number of successive identical nucleotides in reference sequence harbouring a candidate variant
Signal shift (RBI)	≥ 0.25	≥ 0.25	Deviation between predicted and observed flow signal

Relative read quality (MLLD)	≤ 5	≤ 5	Difference in mean log likelihood that the reads covering a candidate variant site are in support of either a variant call or a reference allele (based on read mapping quality and base quality scores)

Additional variant filtering was conducted according to threshold values published in a benchmarking analysis conducted by Damiati *et al.* 2016 [45]. Briefly, variants generated from the NA12878 HapMap sample using the Ion Ampliseq™ library preparation pipeline and TVC 5.2 (NA12878-Ampliseq) were assessed for accuracy against the gold standard variants from the NA12878 consensus sequences obtained from the Genome in a Bottle (GIAB) consortium (NA12878-GIAB) [46]. The proportion of true positive variants (variants and corresponding genotypes observed in NA12878-Ampliseq that *concordantly* matched NA12878-GIAB) and false positive variants (variants and corresponding genotypes observed in NA12878-Ampliseq that *discordantly* matched NA12878-GIAB) were evaluated based on 11 TVC filtering parameters. A total of 5/11 filtering parameters that best differentiated true positive from false positive variants were chosen for both SNVs and INDELs (Table 2.3). The stringency of the filtering thresholds were demarcated as low, medium, and high which corresponded to 90, 95, and 99% of true positive calls retained within the in NA12878-Ampliseq variant call set, respectively.

Table 2.3: Damiati variant filtering thresholds for SNVs and INDELs at 3 stringency levels (adapted from Damiati *et al.* 2016)

		SNV filtering parameters				
Stringency	**% true positives retained**	**Genotype quality**	**Strand bias**	**Variant quality**	**Flow evaluator alternate allele counts †**	**Flow evaluator depth of coverage †**
Low	99	≤ 5	≥ 0.90	20	≤ 2	≤ 6
Medium	95	≤ 8	≥ 0.70	20	≤ 2	≤ 6
High	90	≤ 20	≥ 0.60	30	≤ 2	≤ 10
		INDEL filtering parameters				
Stringency	**% true positives retained**	**Genotype quality**	**Length of homopolymer stretch**	**Variant quality**	**Flow evaluator alternate allele counts †**	**Flow evaluator depth of coverage †**
Low	99	≤ 5	≥ 6	≤ 20	≤ 4	≤ 10
Medium	95	≤ 8	≥ 5	≤ 30	≤ 4	≤ 20
High	90	≤ 20	≥ 4	≤ 40	≤ 4	≤ 25

† Flow evaluator filtering parameters represent values predicted according to flow signal information.

2.4.8 Variant annotation

All variants passing filtering QC underwent gene based annotation using the ANNOVAR *geneanno* pipeline with the refGene database [47]. Specifically, variants were classified to 1 of 9 genomic regions (Table 2.4): 1) exonic, 2) splicing, 3) non-coding RNA (ncRNA), 4) UTR5, 5) UTR3, 6) intronic, 7) upstream, 8) downstream, or 9) intergenic and annotated to the gene harbouring the variant in question. Variants classified as exonic were further classified according to 8 mutation types (Table 2.5): 1) frameshift insertion, 2) frameshift

deletion, 3) stopgain, 4) stoploss, 5) nonframeshift insertion, 6) nonframeshift deletion, 7) nonsynonymous SNV, 8) synonymous SNV. Nonsynonymous SNVs were annotated with 5 *in silico* pathogenicity algorithms to assess whether they conferred deleterious/damaging effects on protein function using version 3.0 of the dbNSFP database (dbNSFP v.3.0) [48] (Table 2.6): 1) Sorting Intolerant From Tolerant (SIFT), 2) Polymorphism Phenotyping 2 HumDiv (Polyphen-2-HDIV), 3) Polymorphism Phenotyping 2 HumVar (Polyphen-2-HVAR), 4) Combined Annotation Dependent depletion (CADD), and 5) Mendelian Clinically Applicable Pathogenicity (M-CAP) [49]. Lastly, variants were annotated according to their corresponding allele frequencies and counts in major external exome sequencing databases using in-house shell scripts. These databases included the 1000 Genomes Phase 3 Project (1KGP3) [50], NHLBI GO Exome Sequencing Project (NHLBI GO ESP) 6500 [51], and version 0.3 of the Exome Aggregation Consortium (ExAC v0.3) [20,21].

Table 2.4: Genomic region annotation descriptions

Mutation type	Short description
Exonic	Sequence within the coding region (exon) of a gene
Splicing	2-bp sequence corresponding to splice donor/acceptor sites
ncNRA	Sequence corresponding to an un-translated, functional RNA molecule
UTR5	Sequence corresponding to the untranslated region on the 5’ end of a gene
UTR3	Sequence corresponding to the untranslated region on the 3’ end of a gene
Intronic	Sequence within the non-coding region of a gene
Upstream	Sequence within 1 kb upstream of transcription start site
Downstream	Sequence within 1 kb downstream of transcription end site

Intergenic	Sequence corresponding to non-coding regions between genes

Table 2.5: Mutation type descriptions

Mutation type	Short description
Frameshift insertion	Nucleotide insertion that alters the reading frame of the mRNA transcript
Frameshift deletion	Nucleotide deletion that alters the reading frame of the mRNA transcript
Stopgain	SNV that results in a premature stop codon
Stoploss	SNV that results in loss of a native stop codon
Nonframeshift insertion	Nucleotide insertion that maintains the reading frame of the mRNA transcript
Nonframeshift deletion	Nucleotide deletion that maintains the reading frame of the mRNA transcript
Nonsynonymous SNV	SNV that results in an amino acid substitution
Synonymous SNV	SNV that results in maintenance of the reference amino acid (i.e. silent mutation)

Table 2.6: *In silico* pathogenicity score descriptions

***In silico* pathogenicity algorithm**	**Mutation types annotated**	**Pathogenicity prediction**
SIFT	Nonsynonymous SNV	Amino acid conservation at the variant site is assessed by comparing sequence homology to closely related species. Amino acid substitutions within highly conserved proteins or protein domains are deemed more deleterious.

Polyphen-2-HDIV	Nonsynonymous SNV	A Naïve Bayes classifier trained using supervised machine learning is trained using variants that are known to be causative for Mendelian diseases as curated in in the UniprotKB database as the diseased set. Polymorphic variants in related species that are known not to induce a functional effect as curated in the UniprotKB database are used as the benign set.
Polyphen-2-HVAR	Nonsynonymous SNV	A Naïve Bayes classifier trained using supervised machine learning is trained using all known disease-causing variants as curated in in the UniprotKB database as the diseased set. Variants classified as common (MAF >1%) with no known involvement in disease onset were used as the benign set.
CADD	Nonsynonymous SNV, nonframeshift insertion/deletion, frameshift insertion/deletion, intronic, intergenic	A support vector machine with linear kernel based on 63 annotations is trained on ~14 million observed human-chimpanzee allelic variations with an allele frequency > 95% in the human genome (i.e. fixed alleles) as the diseased set compared to ~14 million simulated variants as the benign set.
M-CAP	Nonsynonymous SNV	A gradient boosting tree classifier trained using known rare, pathogenic missense variants in the Database of Human Genetic Data (HGMD) as the disease set and rare, missense variants in ExAC as the benign set.

2.4.9 Sex check

Genetic sex was determined by calculating the heterozygous to non-reference homozygous ratios for X-chromosome variants using an in-house shell script. Ratios were subsequently plotted to establish genetically defined male/female clusters. Samples demonstrating inconsistency between reported and genetically determined sex were flagged.

2.4.10 Ethnicity check

Variants used for ethnicity check were comprised of low-frequency and common SNVs (MAF>0.01) with call-rates > 99%. All variants also underwent LD-pruning using plink v1.9 [52] with a window size of 100, window shift of 50, and an r^2 threshold of 0.2. Eigenvectors corresponding to the first 20 principle components were determined using plink v1.9 [52]. The first 2 principle components were used as dependent and independent variables identify ethnic clustering. Eigenvectors for the same principle components were also calculated for reference ethnicities from HapMap 3 which were used to identify discrepancies in reported versus genetic ethnicity. Reference HapMap data was obtained from the HapMap ftp repository: ftp://ftp.ncbi.nlm.nih.gov/hapmap/genotypes/hapmap3/.

2.4.11 Genotypic concordance

All sequenced individuals were genotyped on the HumanCoreExome Beadchip (Illumina). Sequence and genotype data was evaluated for overall genotypic concordance using the GATK GenotypeConcordance tool [44]. Overall concordance was evaluated over the intersection of sites in sequencing and genotype data and was calculated as the proportion of total genotypes that were concordant (i.e. N concordant genotypes / total genotypes).

2.4.12 Sequencing quality metrics

Per-base depth of coverage values were computed using the GATK DepthOfCoverage tool [44] across Ion Ampliseq™ Exome target regions. In house shell scripts were subsequently used to calculate 3 coverage-based metrics to assess the quality of a given exome: 1) mean depth of coverage across target bases, 2) proportion of target bases covered by at least 20 reads (% > 20X coverage), and 3) the proportion of target bases covered by at least 0.2% of the mean depth of coverage (coverage uniformity). Samples that were lower bound outliers (< Q1-1.5xIQR) in any of these metrics after initial sequencing were re-sequenced and combined with their previous exome in order to achieve higher quality. Additionally, any samples achieving < 75% 20X coverage were also re-sequenced and combined with their previous exome regardless if they were outliers in this category or of their performance in the other coverage-based sequencing metrics. 2 additional metrics were used to gauge sequencing quality independent of coverage: 1) proportion of reads mapped to target regions (% on-target reads), and 2) proportion of bases with a phred scaled quality of at least 20 (% Q20 bases). Lower-bound outliers in these 2 categories were re-sequenced, but kept independent of their previous exome. Both non-coverage-based metrics were obtained directly from the ion torrent browser. All metrics were stratified according to sequencing platform (Ion Proton™ or Ion S5XL™).

2.4.13 Statistical analysis

All statistical computations including means and standard deviations (SD) for clinical and sequencing data were calculated using R version 3.2.2 unless otherwise stated. Differences

in sequencing quality metrics across sequencing platform and exome library preparation strategies were assessed using multiple linear regression to account for covariates. Differences in variant counts after implementation of Diamati filtering criteria were determined with either student's t-test (parametric) or wilcoxin rank sum test (non-parametric) depending on if comparison sets were normally distributed. Normality was assessed using the shapiro-wilk test for normality. Figures were generated using the ggplot2 package. All data in the form $x +/- y$ represents $mean +/- SD$, unless otherwise stated.

2.4.14 Induced pluripotent stem cell workflow

For the stem cell arm of DECODE, patient-specific blood cells will be reprogrammed into induced pleuripotent stem cells using ectopic expression of a cocktail of transcription factors such as OCT4, SOX2, and Nanog. These cells will be expanded *in vitro* and differentiated into appropriate lineages. Throughout the differentiation process, the emergence of a highly proliferative, immature cell type (progenitors) will be monitored using a reporter system and proliferation assays (i.e. BrDU/Ki-67 growth curves). The derived cells will then be characterized phenotypically and functionally through the use of lineage-specific markers and gene expression (RNA-seq), respectively.

2.5 Results

2.5.1 Clinical features of the DECODE cohort

The DECODE study population included 40 males and 15 females with mean age (at time of CAD diagnosis) of 35.8 and 39.6, respectively (Figure 2.2 and Table 2.7). Most participants presented with severe coronary disease with 83% having had a NSTEMI or STEMI and 58% with significant occlusion in multiple coronary vessels (Table 2.7). Additionally, 62% of participants declared positive first-degree family history of cardiovascular disease (Table 2.7) including MI (91%), angina (5%), stroke (2%), or peripheral artery disease (2%). Prevalence of CAD risk factors included 46% of participants being previous or current smokers, 44% having hypertension, 9% having a BMI > 40 kg/m^2 (mean 33 kg/m^2), and 17% having non-insulin dependent type II diabetes mellitus (Table 2.5). Median LDL-C and total cholesterol were 3.2 mmol/L and 4.7 mmol/L, respectively with 20% of participants on statin treatment at time of angiography (Table 2.8).

Table 2.7: Summary of clinical features for the DECODE cohort (n=55). Median values are provided for continuous variables.

Clinical feature	Mean +/- SD (median) or proportion
Sex	
Male (%)	73
Age at diagnosis	
All	36.7 +/- 4.3 (37.5)
Male	35.6 +/- 3.9 (37.0)
Female	39.7 +/- 4.1 (41.0)
Clinical outcome (%)	
Stable CAD	17
NSTEMI	35
STEMI	48
No. of diseased coronary vessels (%)	
1	40
2	34
3	26
First-degree family history of CVD (%)	62
Smoking (%)	46
Hypertension (%)	44
Non-insulin dependent diabetes mellitus (%)	17
BMI > 40 kg/m^2 (%)	9
BMI (kg/m^2)	32.6 +/- 7.4 (31.5)

Table 2.8: Summary of lipid panel measurements for the DECODE cohort (n=55). Median values are provided for continuous variables.

Lipid feature	Mean +/- SD (median) or proportion
LDL-C (mmol/L)	3.26 +/- 1.0 (3.2)
HDL-C (mmol/L)	0.937 +/- 0.18 (0.92)
Total cholesterol (mmol/L)	4.85 +/- 1.5 (4.7)
Triglycerides (mmol/L)	2.31 +/- 2.11 (1.8)
Statin treatment (%)	20

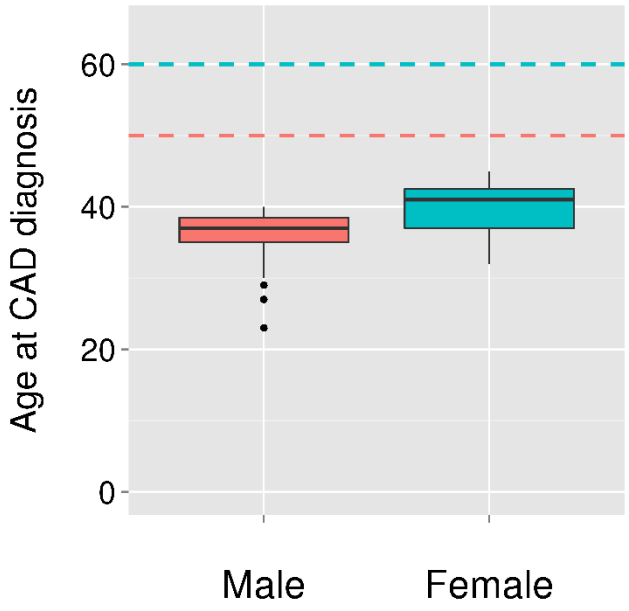


Figure 2.2: Age distribution of males and females in the DECODE cohort. Dashed lines represent the traditional upper-bound age cutoffs for males and females used in most epidemiological studies to classify CAD as early-onset.

2.5.2 Ethnic composition of DECODE cohort

DECODE is predominantly composed of individuals with European ancestry (78%). However, there remainder of the population is ethnically diverse with representatives of multiple ancestries (2 South Asian (6%), 1 African American (2%), 1 Arab (2%), 1 East Asian (2%), 2 European-South Asian Mixed (4%), 3 European-Native Mixed (6%), 1 Latino-Egyptian Mixed (2%)) (Figure 2.3).

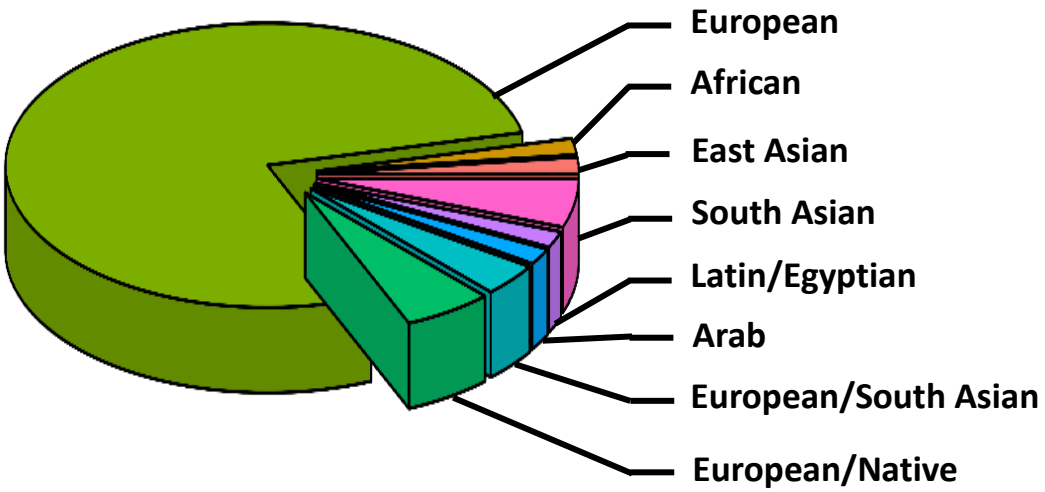


Figure 2.3: Ethnic distribution of DECODE cohort.

2.5.3 Sequencing quality control

Among 6 sequencing based quality metrics (4 coverage-based, 2 non-coverage-based), higher mean depth of coverage was found to associate with lower plexity (i.e. 2 plex chip loading) (P = 0.003; beta = 51.6) after adjustment for sequencing platform and template preparation strategies. Additionally, higher coverage uniformity was associated with template preparation on the IonChef™ after adjustment for sequencing platform and

plexity ($P = 1.4x10^{-4}$; beta = 5.6). Across DECODE exomes, % 20X coverage and both non-coverage-based metrics (% on-target reads and % Q20 bases) demonstrated consistent performance across all strata and were not found to be associated with sequencing platform, plexity, or template preparation strategy ($P>0.05$).

In an additional ~230 non-DECODE exomes that were also run through the IonAmpliseq™ pipeline, the associations for higher mean depth with 2plex chip loading and higher coverage uniformity with IonChef™ template preparation strategy were validated ($P = 3.1x10^{-15}$; beta = 29.5 and $P = 8.8x10^{-12}$; beta = 5.1, respectively). However, higher % 20X coverage in non-DECODE exomes was found to associate with both 2plex chip loading and template preparation on the IonChef™ ($P = 0.007$; beta = 3.7 and $P = 0.005$; beta = 5.4, respectively). For non-coverage-based metrics in non-DECODE exomes, high % on-target reads were modestly (low beta value) associated with 2plex chip loading ($P = 0.0002$; beta = 1.7). No associations were found between % Q20 bases and any strata in non-DECODE exomes. After combining DECODE and non-DECODE exomes, another modest association between % on-target reads and template preparation on the IonChef™ ($P = 5.6x10^{-5}$; beta = 1.2) was identified. All associations individually determined in DECODE and non-DECODE sets were maintained in DECODE + non-DECODE exomes.

Table 2.9: Summary (mean +/- SD) of sequencing quality metrics for DECODE cohort (n=55) stratified by sequencing platform, plexity, and use of Ion Chef™ or Ion OneTouch™ for template preparation and chip loading. Orange and blue shading for coverage-based and non-coverage based parameters, respectively.

Sequencing metric	Ion Proton™	Ion S5XL™	3 plex	2 plex	Ion Chef™	Ion OneTouch™
Mean depth of coverage	75.3 (+/- 21.6)	110.7 (+/- 29.2)	71.7 (+/- 30.1)	110.9 (+/- 15.5)	106.1 (+/- 30.5)	75.4 (+/- 20.0)
% 20X coverage	83.7 (+/- 6.1)	93.7 (+/- 4.5)	83.1 (+/- 5.8)	93.1 (+/- 4.9)	93.0 (+/- 5.0)	83.2 (+/- 4.9)
Coverage uniformity	88.5 (+/- 2.8)	93.6 (+/- 2.4)	88.5 (+/- 3.4)	92.9 (+/- 2.5)	93.6 (+/- 2.2)	84.0 (+/- 2.3)
% on-target bases	94.9 (+/- 0.8)	93.3 (+/- 1.6)	95.0 (+/-0.8)	93.4 (+/-1.5)	93.5 (+/-1.6)	94.9 (+/- 0.7)
% Q20 bases	81.0 (+/- 0.08)	83.9 (+/- 0.01)	81.1 (+/- 0.1)	83.2 (+/- 0.02)	83.6 (+/- 0.01)	80.9 (+/- 0.09)

2.5.4 Variant counts

Default TVC along with low, medium, and high Damiati filters were applied to all reference and variant calls across 55 DECODE exomes as described in section 2.4.6 and Table 2.2. Total variant counts deviated significantly across all stringencies (Figure 2.4 and Table 2.10). Overall, Damiati low, medium, and high stringencies were on average 5, 13, and 35 % lower than default variant filtering settings defined in TVC 5.2, which itself resulted in 42102 (95% confidence interval (CI) 41218-42985) total variants calls per sample (Figure 2.4). The smallest and largest sequential decrease in variant counts was between TVC and low (5%) and medium and high (19%), respectively. All Damiati filters were also applied to ~230 non-DECODE exomes to ensure that the pattern observed for

the DECODE-specific distribution of variant counts was not due to sampling bias. Overall, no significant difference was observed in variant count across each filtering parameter between DECODE and non-DECODE exomes (P>0.05 for TVC, low, medium, and high).

Table 2.10: P-values corresponding to difference in total variant counts between all variant filtering criteria.

TVC vs Damiati low	TVC vs Damiati Medium	TVC vs Damiati High	Damiati Low vs Damiati Medium	Damiati Low vs. Damiati High	Damiati Medium vs Damiati High
0.002	1.23×10^{-9}	2.21×10^{-16}	1.45×10^{-4}	3.2×10^{-16}	3.49×10^{-10}

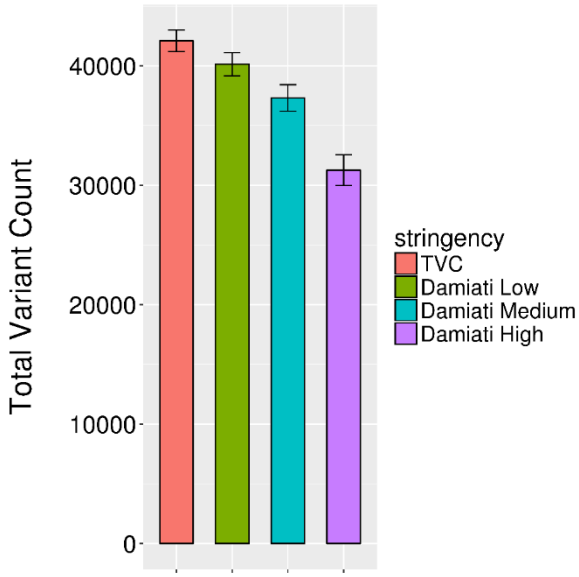


Figure 2.4: Total variant counts for 52 DECODE participants after filtering with default TVC and all Damiati stringencies. Mean values for each filtering criteria are represented by the height of the bars. Error bars depict 95% confidence intervals.

2.5.5 Sex check

The plotting of heterozygous to non-reference homozygous ratios for X-chromosome variants (het:hom-X) resulted in defined clustering of males and females within the DECODE cohort (Figure 2.4). The difference in het:hom-X values between males and females was significant ($P = 6.4 \times 10^{-8}$) with means of 0.28 +/- 0.23 and 1.72 +/- 0.20, respectively. A single individual (DECODE 0014) with reported male sex demonstrated excess heterozygosity on the X-chromosome and was therefore found to group with het:hom-X values of reported female participants (arrow in Figure 2.5). This individual was subsequently flagged and subject to additional quality control to confirm a sex check mismatch (see section 3.3.2).

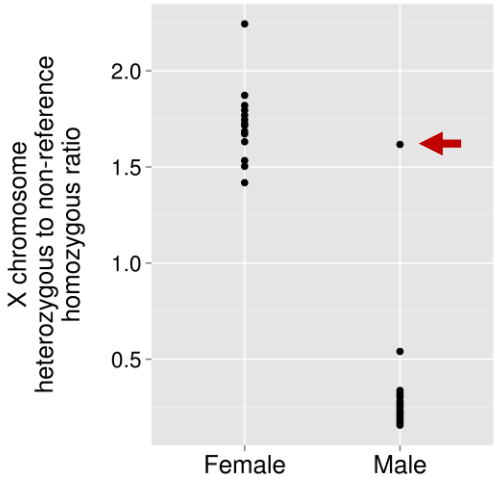


Figure 2.5: Heterozygous to non-reference homozygous ratio for X chromosome variants (het:hom-X) for 55 DECODE participants. Het:hom-X values are stratified by reported sex (male or female). Red arrow represents a reported male sample grouping with the het:hom-X values of reported female participants.

2.5.6 Ethnicity check

No apparent discrepancies between reported and genetic ethnicity were detected.

2.5.7 Genotypic concordance

A single sample (DECODE 0014) was determined to have an overall genotypic concordance of < 90% between exome sequencing and microarray genotyping data. Overall genotypic concordance was 99.4 % (+/- 1.3) across all DECODE participants after excluding the aforementioned sample.

2.6 Discussion

We have herein introduced and described the DECODE study, a multi-arm investigation of 52 EOCAD participants that have underwent whole-exome sequencing in order to identify rare protein-altering variants of high effect with the goal of delineating novel gene-based associations demonstrating both statistical and biological validity with CAD.

We have employed the use EPS to recruit individuals with very early disease in order to enrich for rare causal alleles which empowers downstream statistical approaches to detect rare variant association signals. In fact, several recent analyses have used simulation data to empirically demonstrate substantial increases in statistical power for rare variant associations as sampling from phenotypic distributions becomes increasingly extreme. This is largely evidenced by substantial increases in the MAF of causal variants in extreme populations as the corresponding MAFs in the general population decrease (i.e.

the rarer the causal allele is in the general population, the more enriched it will be in an extreme cohort) [41]. The statistical benefits of EPS have also been demonstrated in application-based studies. Peloso *et al.* 2016 identified that the well-defined association between the rare variants in the ABCA1 gene and HDL-C was substantially more significant when employing extreme sampling versus random sampling ($P = 0.0006$ vs $P = 0.03$) when assessing the rare variants in aggregate [42]. Moreover, the association signal detected in the extreme population required x% fewer participants as compared to the random sample to achieve the same study power. As such, this study effectively illustrates the ability to identify rare variant association signals using EPS which may only demonstrate nominal or no significance with random sampling.

The stringent upper-limit age cutoffs of ≤ 40 for males and ≤ 45 for females establishes DECODE as the youngest cohort to participate in research-based cardiovascular genetics study. Typical upper bound age cutoffs to meet early-onset requirements for CAD are ≤ 50 for males and ≤ 60 for females [2]. Therefore, the 25 and 33% reductions from traditional age cutoffs for DECODE males and females, respectively represents a stringent application of EPS which has yet to be implemented in a research-focused genetic epidemiological studies of early cardiovascular disease. Our study endeavor therefore possesses the ability to identify novel gene-based association signals (driven by the enrichment of rare casual variants within extreme samples) which could not otherwise be detected with more liberal EPS requirements.

The gains in statistical power obtained through extreme phenotypic sampling can be further supplemented by biologically characterizing putative disease-causing alleles.

Thormaehlen *et al.* 2015 restricted analysis to rare missense variants with confirmed functional *in vitro* effects and observed a 10-fold increase in the strength of association between LDLR mutation carrier status and early-MI risk [43]. Therefore, functional modelling resulted in considerable increases in power to detect exome-wide association signals for a given sample size.

Our approach in performing biological validation of candidate variants discovered through exome sequencing meticulously expands upon the methodology employed by the aforementioned flagship study. Specifically, by leveraging the differentiation capacity of induced pleuripotent stem cells, we are able profile rare protein-altering variants in a tissue-specific manner in order to assess their functional effects in the appropriate cellular lineage. In doing this, we are not limited in our capacity to biologically model variants from across exome, which represents a remarkable opportunity to establish novel and sound relationships between gene and phenotype.

Among all DECODE participants, CAD risk factors including smoking, hypertension, insulin-dependent diabetes mellitus, and BMI were modestly prevalent (Table 2.7) and will have to be adjusted for in downstream analyses in order to avoid association signals that are mediated by confounding variables. Median lipid panel measurements were not indicative of cohort-wide dyslipidemia even when taking into account use of cholesterol-lowering medications (Table 2.8 which lends evidence to the variety of molecular phenotypes that can potentially characterize individuals with EOCAD.

We further demonstrate potential sequencing workflow optimization in order to trend toward achieving higher mean depth, % 20X coverage, and coverage uniformity with

IonAmpliseq™ Exomes. Firstly, both mean depth and % 20X coverage can be optimized when multiplexing 2 samples per sequencing chip as opposed to 3. This is largely attributable to more sequencing wells available per-sample which increases the capacity for identical target regions to achieve a high degree of sequence information. This ultimately translates to higher absolute coverage. Secondly, we find that use of the IonChef™ for template preparation results in higher coverage uniformity by streamlining the processes exome library amplification, enrichment and chip loading. The sequential automation of these processes ultimately allows for template sequences to effectively saturate all wells on a sequencing chip which maximizes the proportion of target regions that achieve sequencing read coverage and also minimizes sample to sample variability in coverage uniformity incurred by human error. These findings collectively demonstrate the importance of identifying associations among core sequencing metrics in order minimize preventable loss in data quality.

In order to evaluate the sensitivity of our variant calls, we leveraged the emergence of publicly available consensus sequences have provided the opportunity to conduct benchmarking analyses in order to ascertain accuracy of local variant calling/filtering algorithms for given sequencing chemistries and workflows. Since consensus sequences are devoid of poorly mapped genomic regions (i.e. regions harbouring segmental duplications, short tandem repeats, and copy number variations) and contain highly reliable variants calls that are mutually detected across multiple sequencing chemistries and variant callers, they can serve to modulate the sensitivity and specificity of variant calling/filtering procedures based long-term study objectives. The recent work published by Damiati *et al.*

2016 was able to identify 5 variant filtering metrics for SNV and INDELs called from exomes generated using the IonAmpliseq™ Exome library preparation strategy and TVC [45]. After employing these filtering criteria at 3 stringency levels (low, medium, and high), we observed significant decreases in total variant counts (Table 2.3 & Figure 2.4). While gains in positive-predictive value are expected with increased filtering stringency, we are also limiting our sensitivity to detect true gene-based associations as we increase variant filtering stringency due to the inflation of false negative variant calls. Therefore, it is essential implement higher variant filtering stringencies with caution in order to avoid missing true associations. Results pertaining to variant calling benchmarking using in-house sequencing workflows are described in Chapter 6.

Finally, we used 3 sample level quality control procedures (sex check, ethnicity check, and genotypic concordance) that are adept at detecting aberrations during library and template preparation such as sample swaps and sample admixtures (i.e. sample contamination due to the mixing of DNA from two or more samples). Sex check revealed a discrepancy in a sample that was reported male, but demonstrated excess heterozygosity for X-chromosome variants which is indicative of female sex. This sample also had low overall genotypic concordance, but not to an extent that warrants evidence of sample swap. Therefore, these results collectively point to a case sample admixture which is further supported by evidence provided in section 3.3.2 where this sample was also found to exhibit excess autosomal heterozygosity. Although, no further discrepancies were identified among other DECODE participants, it is essential to institute stringent sample-level quality control in order to prevent the inclusion of samples that could lead to false

positive or negative association signals, especially among larger sample sizes where the likelihood of observing samples that fail quality control procedures increases by chance alone.

CHAPTER 3 – Whole Exome Quality Control: Understanding Patterns of Genetic Variation

3.1 Introduction

Over the past five years, exome sequencing has demonstrated a marked increase in research utility by facilitating the discovery of novel genes involved in complex disease susceptibility and protection [2,26,27]. In order to reach sound statistical and biological conclusions on genetic predisposition, it is essential to conduct robust quality control of variant calls at the sample level by leveraging population genetic effects which can assess the quantity and distribution of different variant types. Such effects include (but are not limited to) the proportion of variants previously identified and curated, ratio of heterozygous to homozygous genotypes, and the ratio of transition to transversion mutations. Since these metrics are largely dependent upon population genetic phenomena, the values obtained from locally sequenced samples can be evaluated against those expected due to phenomena such as genetic drift, genetic bottlenecks and purifying selection. More formally, the values derived from local sequences can be directly compared to corresponding statistics obtained from high-quality sequencing datasets (e.g. 1KGP3). This can prove invaluable for detecting quality confounders such as sample contamination, sample relatedness and spurious variant calling which can all critically impact the outcomes of sequencing based epidemiological analyses. These parameters are also useful when assessed in conjunction with sample ethnicity checks since variant types and counts differ significantly between genetically diverse and inbred populations, such as in Africans and Finnish-Europeans, respectively. Therefore, it is good practice to stratify the calculation of these quality control metrics by ethnic composition should there be a sufficient number of samples representative of different ancestries. Overall, conducting

systematic quality control of sequencing data with population-based metrics can provide a comprehensive understanding of the patterns of different classes of genetic variation and can be used to mark and remove samples with significant deviations from expected values.

3.2 Methods

SNVs and INDELs called using TVC 5.2 (manufacturer settings) (fully described in section 2.3.6) for 55 DECODE samples were evaluated through 6 quality metrics ideal for sequencing data: 1) percent concordance with dbSNP version 146 [53], 2) heterozygous to non-reference homozygous ratio, 3) transition to transversion ratio, 4) nonsynonymous SNV to synonymous SNV ratio, 5) frameshift INDEL to non-frameshift INDEL Ratio and 6) singleton counts. Where applicable, these metrics were stratified according to 1) variant curation in dbSNP (known/novel), 2) genomic region (non-coding/coding), and 3) MAF within coding regions (rare coding/common coding). Singleton counts were further stratified according to mutation type (either nonsynonymous or synonymous SNV). DECODE sample variants were annotated with rsIDs from the dbSNP 146 database using SnpEff software. Variants without a corresponding rsID were identified as novel. Variants were annotated to genomic region and mutation type using ANNOVAR *geneanno* pipeline (described in further detail in section 2.3.7) and according to allele frequency using their corresponding MAF in the ExAC v0.3 database. Coding variant were defined as falling within exons or splice donor/acceptor sites and rare variants were defined has having MAF < 1% in all major ExAC ethnicities.

Statistical significance between all stratified pairings for the six population genetic parameters was tested using either a student's t-test (parametric) or wilcoxin rank sum test (non-parametric) depending on whether the data was normally distributed. Normality was tested using the shapiro-wilk test of normality. Due to a small sample size, samples were flagged and marked for exclusion only if they demonstrated extreme deviation from the sampling distribution (less than or greater than 6 times the inter-quartile range (IQR)) after accounting for ethnicity.

3.3 Results and Discussion

3.3.1 dbSNP 146 concordance

dbSNP represents a publically available repository of curated multi-ethnic genetic variation. Currently, version 146 contains ~ 153 million variant sites (91% SNV and 9% INDEL) with more being added to upcoming releases of the database. Concordance with dbSNP allows one to determine what proportion of variants called per sample have previously been curated (i.e. are considered known). A low proportion of known variants is typically indicative of an excess of false positive calls due to poor read alignment or liberal variant QC. However, it is essential to stratify dbSNP concordance according to MAF since rare variants, by virtue of being inherently in-frequent, will be less likely to have been previously reported whereas common variants should be expected to near full concordance. This pattern is observed in the DECODE cohort as novel variants composed a significant portion of rare relative to common variants ($P=3.3 \times 10^{-19}$) with 99.5% of common variants being curated (Figure 3.1B and Table 3.2). Additionally, an increased

proportion of known variants was observed among coding regions ($P=2.0x10^{-11}$) (Figure 3.1A and Table 3.2). This is likely a consequence of the extensive use of whole-exome sequencing strategies in research and clinical-based studies which inherently limits the ability to curate non-coding variants. This observation is in stark contrast to earlier reports (using older versions of dbSNP) claiming that the majority of coding variation is novel. Therefore, it is essential to annotate variants using up to date databases in order to properly assess the distribution of novel and rare alleles across the genome.

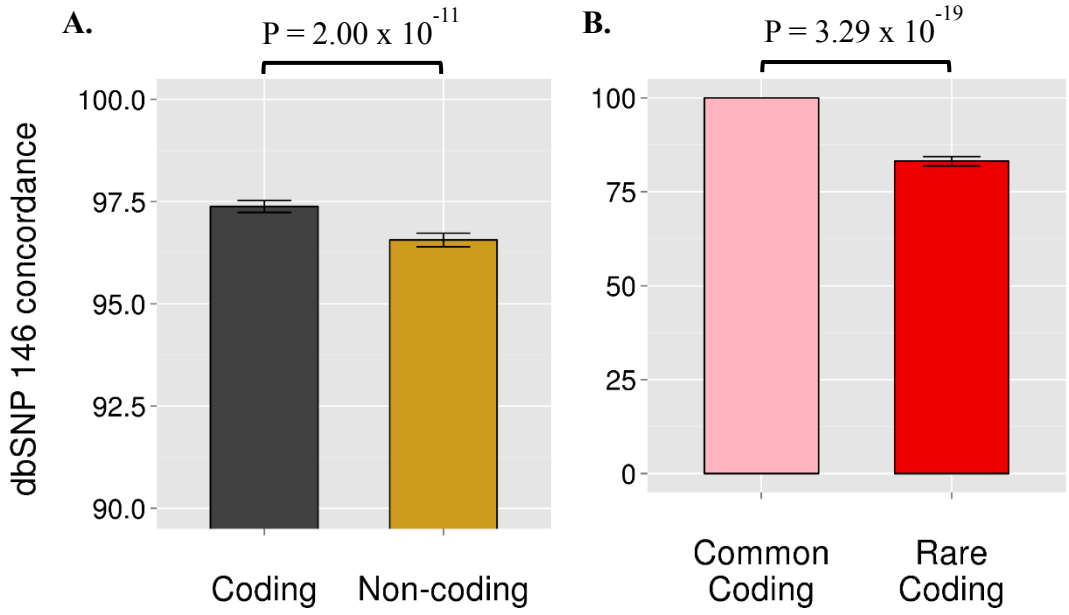


Figure 3.1: Variant concordance with dbSNP 146 in 52 DECODE participants. Percent mean concordance is grouped by **(A)** genomic region (i.e. coding or non-coding) and **(B)** allele frequency in coding regions (common coding or rare coding). Error bars depict 95% confidence intervals. The standard error did not deviate appreciably from the mean for common coding variants.

3.3.2 Heterozygous to non-reference homozygous ratio

The heterozygous to non-reference homozygous (het:hom) ratio is an useful parameter to detect sample admixture caused by multi-sample contamination. Admixed samples will result in excess heterozygosity due to the presence of multiple alleles at candidate non-reference homozygous sites. The genotypes for all variant calls from the autosomal chromosomes were evaluated to assess both genotyping quality and degree of genetic variation within samples, as non-European samples (that are not founder populations) are likely to be more genetically diverse and consequently harbour more heterozygous variation. A single sample (DECODE 0014) of mixed European and South Asian ancestry demonstrated extreme heterozygosity among all autosomal variants ($het:hom_{0014}$ = 3.3; $het:hom_{DECODE}$ (mean +/- SD) = 1.8 +/- 0.1) which is likely a consequence of sample admixture. This sample was consequently flagged for downstream analysis. The remaining upper-bound outliers represented individuals of South Asian and African ancestry (DECODE 0005 and DECODE 0018) and displayed only modest heterozygous inflation. Therefore, these samples were not considered to have excess heterozygosity, especially when considering their ethnicity among a largely European cohort. Novel and rare variants demonstrated significantly higher het:hom ratios compared to known and common variants ($P = 1.5 \times 10^{-9}$ and $P = 6.7 \times 10^{-15}$) (Figure 3.2 and Table 3.2). Homozygous variants (if found to be disease-causing) are predicted to have greater effect on disease risk and often result in more extreme disease phenotypes due to gene dosage effects. Therefore, it is unlikely that individuals will harbour 2 copies of a deleterious variant which explains depletion of homozygous genotypes among novel and

rare alleles. No significant association in het:hom ratio was identified when stratified by coding and non-coding regions (P>0.05).

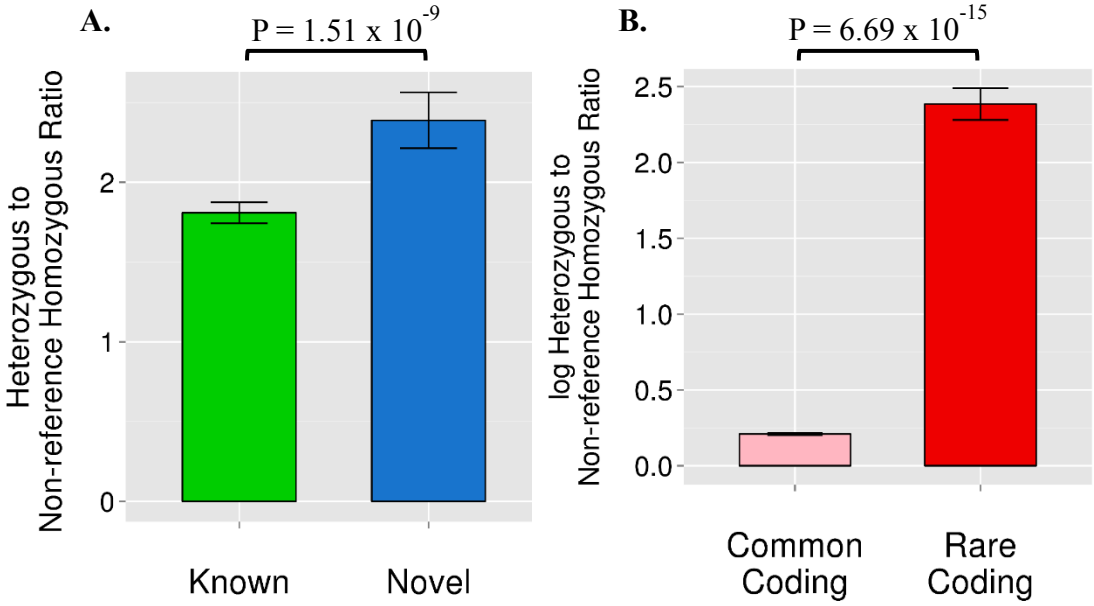


Figure 3.2: Heterozygous to non-reference homozygous ratio for variant genotypes in 52 DECODE participants. Mean het:hom is grouped according to curation (known or novel) **(A)** and allele frequency in coding regions (common coding or rare coding) **(B)**. *Due to substantial differences in het:hom ratio between common and rare coding variants, the y-axis is plotted on a log scale .* Error bars depict 95% confidence intervals.

3.3.3 Transition to transversion ratio

Transitions and transversions are single nucleotide variants that are characterized by purine-purine (A <-> G) or pyrimidine-pyrimidine substitutions (C <-> T) (transitions) and purine-pyrimidine substitutions (A <-> C; A<->T; G<->C; G<->T) (transversions), respectively. It has been consistently shown that transversion variants are enriched amongst protein-altering genetic alterations such as nonsynonymous and nonsense

mutations compared to transitions which tend to result in silent mutations and a more conserved polypeptide sequence [2,18,26,54]. Consequently, transversions are not typically observed as readily as transitions in protein-coding regions. This may also be partially attributable to C to T transitions at CpG dinucleotides. CpG regions are hotspots for epigenetic modification as cytosine nucleotides that become methylated can undergo spontaneous deamination at the 6' carbon to generate thymine. This renders CpG dinucleotides as highly mutable, resulting in a mutation rate that is nearly 10-fold greater than other dinucleotide pairs [55]. As such, it is ideal to calculate non CpG transition to transversion ratios in order to account for preferential mutation bias.

In line with this evidence, it is observed that protein-coding regions are depleted for transversion variants relative to non-coding regions ($P = 1.5 \times 10^{-18}$) (Figure 3.4 and Table 3.2) even after comparing to non CpG transitions. Additionally, there is significant abundance of transversions among novel and rare alleles ($P = 1.9 \times 10^{-64}$ and $P = 1.0 \times 10^{-12}$) (Figure 3.3 and Table 3.2). This supplements evidence that transverions are maintained at low frequency by purifying selection to maintain protein-conservation.

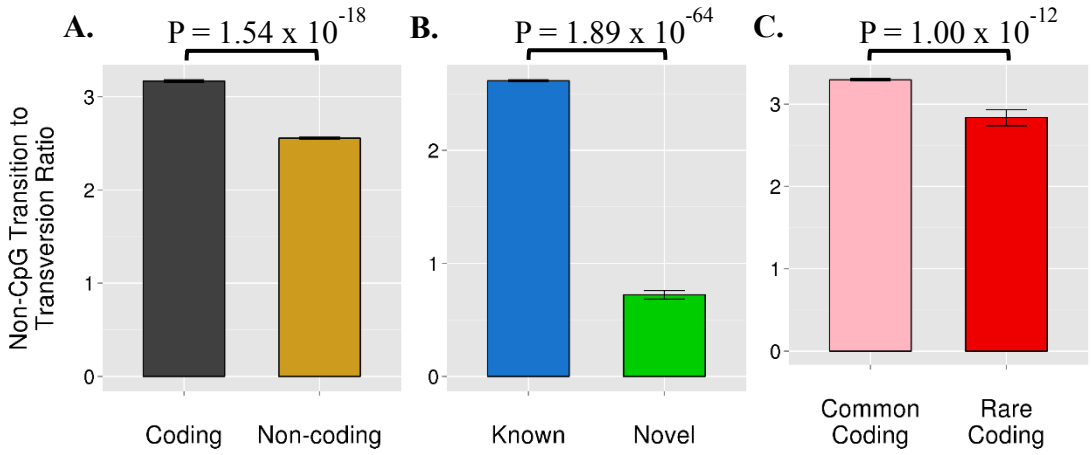


Figure 3.3: Non-CpG transition to transversion ratio for variants in 52 DECODE participants. Mean non-CpG transtion:transversion is grouped according to genomic region (coding or non-coding) **(A)**, curation (known or novel) **(B)** and allele frequency in coding regions (common coding or rare coding) **(C)**. Error bars depict 95% confidence intervals.

3.3.4 Nonsynonymous to synonymous SNV ratio

Nonsynonymous (NS) and (S) mutations are SNVs that result in amino-acid substitutions and maintenance of the reference amino acid, respectively. Novel and rare alleles are enriched for NS mutations ($P = 2.9 \times 10^{-30}$ and $P=7.3 \times 10^{-33}$) (Figure 3.4 and Table 3.2) since variants that alter conserved protein structure are selected against.

3.3.5 Frameshift to non-frameshift insertion/deletion ratio

Frameshift (FS) and non-frameshift (NFS) mutations result from insertions or deletions (INDELs) within the protein-coding regions of the genome. FS INDELs perturb the reading frame of the protein-coding sequence which severely disrupts structure and function of the translated protein and often results in complete loss-of-function for the affected allele. Therefore, FS mutations are strongly selected against and are observed

more frequently among novel and rare alleles ($P = 1.5 \times 10^{-18}$ and $P = 1.9 \times 10^{-12}$) (Figure 3.5 and Table 3.2).

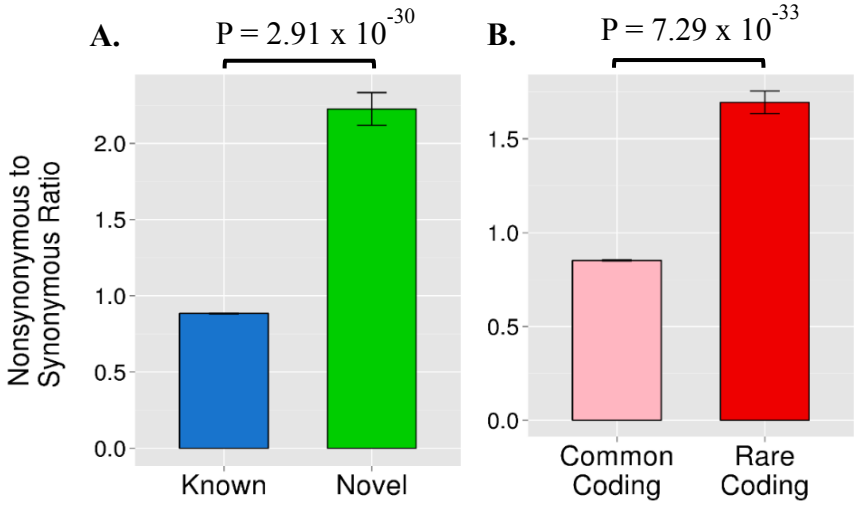


Figure 3.4: Nonsynonymous to synonymous SNV ratio for variants in 52 DECODE participants. Mean NS:S is grouped according to curation (known or novel) **(A)** and allele frequency in coding regions (common coding or rare coding) **(B)**. Error bars depict 95% confidence intervals.

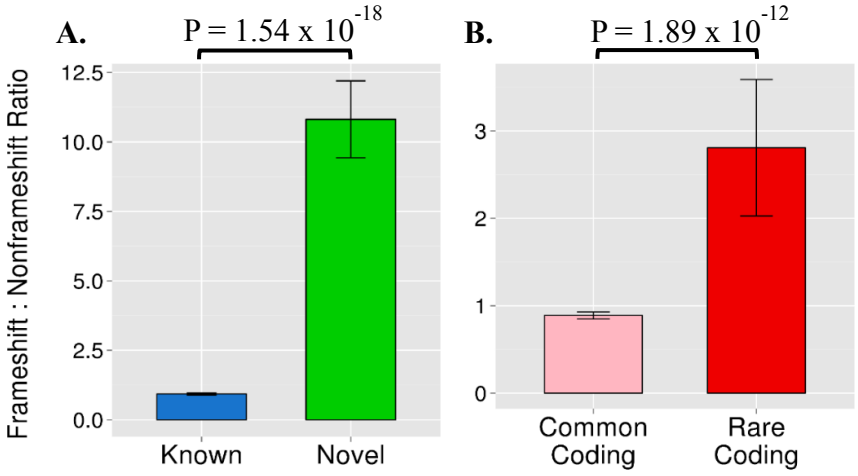


Figure 3.5: Frameshift to non-frameshift INDEL ratio for variants in 52 DECODE participants. Mean FS:NFS is grouped according to curation (known or novel) **(A)** and allele frequency in coding regions (common coding or rare coding) **(B)**. Error bars depict 95% confidence intervals.

3.3.6 Singletons

Singletons are traditionally defined as variants that occur in a single sample from a study cohort in a heterozygous state (i.e. private variants). Similar to het:hom ratio, singleton counts are largely a function of selection pressure, ancestry and sample admixture due to contamination during sequencing preparation. Singletons can also prove useful to mark poorly sequenced or closely related/duplicate samples. Specifically, samples with high kinship will have a depletion in singletons as they are likely to share variants that would otherwise be private. Conversely, samples that contain an excess of singletons are likely to contain an abundance of false positive variants which are unlikely to be observed in other individuals.

When observed in large sequencing cohorts, singletons are preferentially observed amongst novel and protein-altering variants. However, this trend cannot readily be observed in smaller populations as intra-cohort singletons are unlikely to remain private in cohorts with much larger sample sizes (Figure 3.6A). In fact, for 77913 total DECODE singletons, only 2391 (3%) were private after cross-matching with their corresponding allele counts in ExAC (Figure 3.6A and Table 3.1). In contrast, over 45 000 variants found to be private in DECODE were observed to have allele counts greater than 4 in ExAC (Figure 3.6A and Table 3.1) which attests to the necessity of leveraging large publically available databases to classify variants within smaller cohorts. A total of 26330 (33%) DECODE singletons were not observed in ExAC (i.e. DECODE-specific) which either indicates that these sites are false positives or genuine private mutations (Figure 3.6A and Table 3.1). For purposes of this analysis, focus will be kept on DECODE singletons also

found to be private in ExAC in order to confidently establish how singleton mutations are distributed across the genome and identify in which classes of mutations they show enrichment.

Table 3.1: DECODE singletons in the ExAC dataset

Category	DECODE singletons
Present in ExAC	51581
Private in ExAC	1985
DECODE-specific	26330
Total (Present in ExAC + DECODE-specific)	77913

In contrast to findings published in the ExAC flagship paper, singleton mutations are profoundly more enriched among known variants ($P = 1.1x10^{-20}$). In fact a total of only 2 singleton variants were novel across all DECODE participants. This discrepancy between these findings and what is observed in ExAC can largely be attributed to the newer and larger version of dbSNP used to curate variants in this work. Similar to what is reported in ExAC, singletons were observed more frequently among protein-altering variants ($P = 2.4 \times 10^{-9}$) (Figure 3.6B and Table 3.3). This is by virtue of the fact that variants which modify the protein sequence will reduce fitness for survival and will therefore be maintained at lower frequencies in the general population.

Two samples of African and Latin-African ancestry (DEOCDE 0018 and DECODE 62, respectively) showed a high abundance of singleton variants. As previously stated, this is largely driven by ancestry as opposed to sequencing artefacts. Two European samples were observed to have an excessively low singleton count (DECODE 1039 and DECODE 66) and were identified as duplicates. DECODE 1039 was thereafter eliminated from downstream analysis.

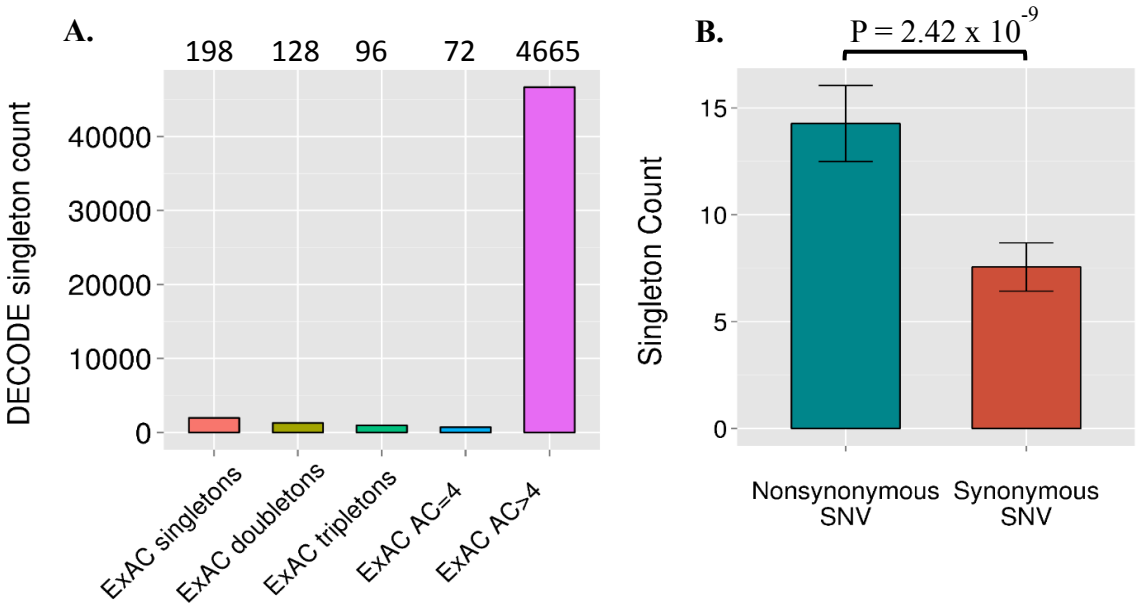


Figure 3.6: Singleton counts observed in 52 DECODE participants. (A) Corresponding ExAC allele counts for DECODE singletons. Mean count for DECODE singletons also observed to be private in ExAC were stratified according to mutation type **(B)** and curation in dbSNP 146 (not shown). Error bars depict 95% confidence intervals.

3.4 Conclusion

Determining the distribution of population genetic metrics across a variety of variant categories can prove invaluable for developing a sound understanding of patterns of genetic variation. Moreover, population genetic metrics calculated within high quality sequencing datasets can be used to as references to gauge the accuracy of variant calls generated in local sequences in order to eliminate individuals that significantly deviate from expected values. This will ultimately serve to establish a clean dataset that will minimize false discoveries in association analyses

Here, we demonstrate patterns of genetic variation consistent with what is expected under the model of purifying selection by evaluating six population genetic effects in 55 ethnically diverse samples that have been exome sequenced. Specifically, variants which are kept at rare frequencies in the population due to having deleterious effects on fitness for survival are observed to be enriched for heterozygous genotypes and for mutation types that **severely** alter the structure of translated protein products (i.e. nonsynonymous SNVs and frameshift INDELs). This pattern supports the undertaking of rare variant association analyses in order to identify alleles that have an impactful effect on disease phenotype and account for heritably that has yet to be explained by epidemiological studies focusing on common variants. We also establish that coding sequences preferentially harbour non-CpG transition mutations, indicating that these regions are under high evolutionary constraint. Therefore, exome sequencing will serve as an ideal approach to maximize the sensitivity of detecting coding variants and increase the power of downstream rare variant association testing.

Furthermore, we demonstrate that applying population genetic features such as het:hom ratio and singleton counts to the raw sequences of all DECODE study participants allowed for the identification and subsequent removal of an admixed and duplicate sample (DECODE 0014 and 1039, respectively). Inclusion of these samples in downstream analysis may have skewed association signals due to inaccurate genotyping and inflated allele counts.

Table 3.2: Mean (95% CI) values for six population genetic metrics across n = 52 DECODE individuals

Metric	Coding variants	Non-coding variants	Known variants	Novel variants	Rare coding variants	Common coding variants
dbSNP 146 concordance	97.4 (97.2-97.5)	96.6 (96.4-96.7)	NA	NA	83.1 Φ (81.9-84.3)	99.9 (99.9-99.9)
het: hom ratio	1.78 (1.75-1.81)	1.80 (1.77-1.84)	1.78 † (1.75-1.81)	2.39 (2.22-2.56)	306.42 Φ (258.26-354.58)	1.62 (1.59-1.65)
Non CpG Transition: Transversion Ratio	3.17 ¥ (3.16-3.18)	2.56 (2.55-2.57)	2.62 † (2.61-2.63)	0.72 (0.69-0.76)	2.84 Φ (2.73-2.93)	3.29 (3.18-3.31)
NS:S Ratio	NA	NA	0.884 † (0.881-0.886)	2.23 (2.12-2.33)	1.69 Φ (1.63-1.75)	0.852 (0.849-0.854)
FS:NFS Ratio	NA	NA	0.927 † (0.885-0.969)	10.8 (9.46-12.2)	2.81 Φ (2.04-3.56)	0.888 (0.850-0.927)
Singletons	24.2 ¥ (21.5-27.0)	13.9 (12.3-15.5)	38.1 † (34.2-42.1)	0.301 (0.212-0.415)	NA	NA

¥ $p < 0.00001$ (coding vs. non-coding); † $p < 0.00001$ (known vs. novel); Φ $p < 0.00001$ (rare coding vs. common coding)

Table 3.3: Mean (95% CI) singleton counts stratified by mutation type for n = 52 DECODE samples

Nonsynonymous SNV	Synonymous SNV
14.3 (12.5-16.0) Ψ	7.56 (6.45-8.66)

Ψ $p < 0.00001$ (nonsynonymous SNV vs. synonymous SNV)

CHAPTER 4 – Assessing the Prevalence of Mendelian Dyslipidemias in an Early CAD Population and Considerations for Clinical Intervention

4.1 Introduction

Exome sequencing has been successfully applied in population-based research settings to identify high impact variants that aggregate amongst diseased cases or healthy controls to confer risk and protective effects, respectively [2,26,30]. More recently, however, exome sequencing has demonstrated tremendous promise in clinical settings for molecular diagnosis of rare diseases that are suspected to have an underlying genetic etiology. The utility of clinical exome sequencing (CES) can effectively be quantified by the proportion of sequenced cases with successful identification of a disease-causing variant within a candidate loci for a given clinical outcome (i.e. the diagnostic yield). Recent reports have demonstrated the efficacy of CES primarily in the context of neurological disorders for both trio (parent-child) and proband (single sample) cases. Whole exome sequencing conducted by Yang *et al.* 2013 [56] and Lee *et al.* 2014 [57] determined an overall diagnostic yield of ~25% for individuals characterized by genetically heterogeneous neurological disorders (e.g. developmental delay, autism, intellectual disability) which is superior to other genetic tests such as karyotype analysis and chromosomal microarray analysis (5-20%) in individuals with similar phenotypic composition. Moreover, the diagnostic yield observed in the Yang et al. 2013 study was achieved from unselected cases (i.e. individuals characterized by a wide range of phenotypic abnormalities) which further demonstrates versatility of CES at detecting causal variants for a spectrum of clinical presentations.

Although CES has proved invaluable for disease diagnosis, its long-term utility for enhancing prognostic outcomes remains to be determined. Therefore, shifting focus to clinically actionable Mendelian disorders may provide an avenue towards assessing the

implications of CES on informing critical lifestyle changes and pharmacological interventions. Familial hypercholesterolemia (FH) is the most common genetic Mendelian disease and an important risk factor for CAD [32,58]. However, the frequency of FH in early CAD patients remains controversial and clinical criteria for FH screening are not adapted to early CAD patients. For instance, ongoing treatment with lipid lowering agents is expected to be the norm in this population such that cholesterol criteria might not apply in FH diagnosis. The DECODE study aims to fill this knowledge gap through systematic genetic screening of young CAD patient for FH-causing mutations using whole-exome sequencing and a semi-automated bioinformatics pipeline tailored for the detection of disease-causing mutations in known FH genes. For comprehensiveness, we also aim to characterize the prevalence of other Mendelian dyslipidemias such as familial combined hyperlipidemia (FCH), sitosterolemia and hypertriglyceridemia in the DECODE cohort.

4.2 Methods

4.2.1 DECODE study population

The DECODE cohort was used to assess the enrichment of FH in an early CAD population. Details on the population are provided in section 2.3.1.

4.2.2 MIGen and CHARGE consortia

A total of 8577 CAD-free individuals from the Myocardial Infarction Genetics (MIGen) consortium [59] (encompassing 7 CAD case-control cohorts) and 11908 individuals from the Cohorts for Heart and Aging Research in Genetic Epidemiology (CHARGE) consortium

[60] (encompassing 5 prospective population-based cohorts) were used as the reference population. Information on variant counts meeting specified pathogenicity criteria were mined from data published by Khera *et al.* 2016 [61].

4.2.3 Exome sequencing, variant calling, and variant annotation

Exome sequencing, variant calling, and variant annotation were all conducted as described in sections 2.4.5-8.

4.2.4 Variant filtering and phenotype matching

Variants leading to a change in the primary amino acid sequence (i.e. protein-altering) that had MAF < 0.05 (i.e. rare) in all major ethnicities of external databases (described in section 2.3.7) were retained. These variants were further filtered at a MAF < 0.05 within an in-house dataset consisting of 230 processed using the same protocol to identify potential sequencing artefacts. Analysis was further restricted to 24 genes (defined by the Western Database of Lipid Variants (WDLV)) [62] in which mutations are known to cause monogenic dyslipidemias and confer substantial risk for developing early CAD. Rare, protein altering variants present in any of the 24 genes were further assigned a pathogenicity ranking using the ClinVar database (NCBI) [63]. Variants identified as "likely pathogenic" or "pathogenic" were concluded to be causal for early CAD, whereas "variants of uncertain significance" (VUS), variants with conflicting interpretations, or non-curated variants were manually interrogated using the InterVar software [64] in order to facilitate a standardized interpretation on variant pathogenicity. Variants unambiguously annotated as "likely benign" or "benign" after inspection in InterVar were not further considered.

Putative disease-causing variants ("pathogenic", "likely pathogenic", or VUS) were retained and confirmed with Sanger sequencing.

Individuals harbouring the variant of interest were deeply phenotyped in parallel to variant classification to determine genotype-phenotype concordance. Briefly lipid panel metrics including LDL-C, HDL-C, total cholesterol, apolipoprotein B to apolipoprotein A1 ratio (apoB/apoA1), triglycerides along with history of treatment regimens were used to assign each variant carrier to a dyslipidemia class according to the Frederickson Classification of Lipid Disorders [65]. Lastly, family history information was used to determine potential co-segregation of the variant of interest with early CAD/MI. Variant and phenotype level data was then merged to assess causality. A schematic describing this process is provided (Figure 4.1).

4.2.5 Detailed coverage calculations

The % 20X coverage for the 24 genes described in WDLV was determined by first generating per-base coverage metrics for the coding intervals of each gene using the GATK DepthofCoverage tool. Using in-house shell scripts, per-base coverage values were used to calculate % 20X coverage by assigning each base with a coverage ≥ 20X to value of 100 and a coverage of < 20X to a value of zero and subsequently computing the mean proportion of bases ≥ 20X per exon.

4.2.6 Sanger sequencing preparation.

Sample DNA harbouring a variant(s) of interest was amplified using a standard PCR protocol. A 20 uL PCR reactions containing 1-5 ng/ul starting DNA, 50 um forward and

reverse primer (Sigma Aldrich), 1X PCR Buffer II (Life Technologies), 3mM Magnesium Chloride ($MgCl_2$) (Life Technologies), 0.2 mM dNTP mix (deoxyribonucleotides) (Fermentas), and 0.1 U/ul AmpliTaq® Gold Polymerase was used for all amplification runs. DNA was amplified using a C1000TM thermocyler (BioRad) according to cycling times and temperatures summarized in Table 4.1. PCR amplicons were purified using the MinElute® Reaction Cleanup Kit with manufacturer protocol and were quantified to 1ng/ul per 100 bases using the Qubit™ dsDNA High Sensitivty Assay Kit and the Qubit™ 2.0 fluormeter (ThermoFisher). 5 ul of DNA along with 5 ul of 1uM forward and reverse primer was sent to the MOBIX laboratory at McMaster Univesity for Sanger sequencing. Summary information for the primers used in this chapter, including sequence and amplicon length, are provided in Supplementary Table 1 (Table S1).

Table 4.1: PCR run parameters for Sanger sequencing preparation

Stage	Temperature (°C)	Time (min)	Cycles
Polymerase activation	95	3:00	1
Denaturation	95	0:30	
Annealing	60 → 50 (gradient)	0:30	34
Extension	72	0:30	
Final Extension	72	3:00	1

4.2.7 Nomenclature

Clinically relevant variants will be designated according coding DNA reference sequence and protein reference sequence as described by Human Genome Variation Society (HGVS) [66]. Coding DNA nucleotide substitutions will be prefixed by "c." followed by the nucleotide's position on the coding sequence and the nucleotide substitution (e.g. c.43G>A

refers to a A → C substitution at position 43 of the coding sequence for a given gene). Coding DNA nucleotide deletions will be prefixed with "c." followed by the positions and sequence of the deleted nucleotides (e.g. c.2397_2400delCGTC refers to a 4 base pair deletion of 'CGTG' nucleotide sequence from position 2397-2400 in the coding sequence of a given gene).

Protein-level substitutions will be prefixed by "p." followed by the reference amino acid, amino acid position in the primary protein sequence, and the alternate amino acid (e.g. p.Asp4Tyr refers to an aspartic acid to tyrosine substitution at position 4 of the primary protein sequence). Protein-level deletions are prefixed by "p." followed by the first and last amino acid(s) and amino acid position(s) that are deleted (e.g. p.Val800_Leu802del indicates a deletion of amino acids at positions 800, 801, and 802 of the primary amino acid sequence where position 800 encoded a Valine and position 802 encoded a Leucine).

It is important to note that the descriptions provided above represent a means to facilitate the understanding of results presented in this chapter and are not inclusive of all the mutation types for which there are nomenclature standards.

4.2.8 Clinical databases and software tools

The ClinVar 2016 database along with the InterVar software too were used to assess the pathogenicity of clinically actionable variants.

The ClinVar database is a publically available archive which curates the pathogenicity conferred by sequence variants on human disease. The degree of pathogenicity is based on submission of research and/or clinical reports outlining the effect of a variant at the cellular,

individual, or population level. The confidence of the pathogenicity ranking ascribed to a given variant is dependent on the depth and validity of supporting information.

InterVar is a bioinformatics software tool that facilitates standardized clinical interpretation of putative disease-causing variants by generating automated interpretation of 18/28 guidelines published by the American College of Medical Genetics (ACMG) and Association for Molecular Pathology (AMP). The additional 10 guidelines are required to be manually curated by the user based on specific attributes for the clinical case and variant in question (e.g. family history of disease under evaluation, established functional effect of variant *in vitro* or *in vivo* studies).

The M-CAP in silico pathogenicity score has demonstrated tremendous efficacy at correctly classifying known pathogenic variants and was used to supplement pathogenicity rankings obtained through ClinVar and InterVar.

4.2.9 Statistical analysis

To test the significance between early CAD status and frequency of FH-mutations, *P* values were calculated using a 2-tailed Fisher's exact test. Odds ratios and 95% confidence intervals were calculated directly from the contingency matrix. A p value of < 0.05 was used as a significance threshold. The 95% confidence interval for proportions was calculated according to the modified Wald method. All data in the form x +/- y represents $mean$ +/- SD unless otherwise stated. All statistical computations were conducted in R version 3.2.2 unless otherwise stated.

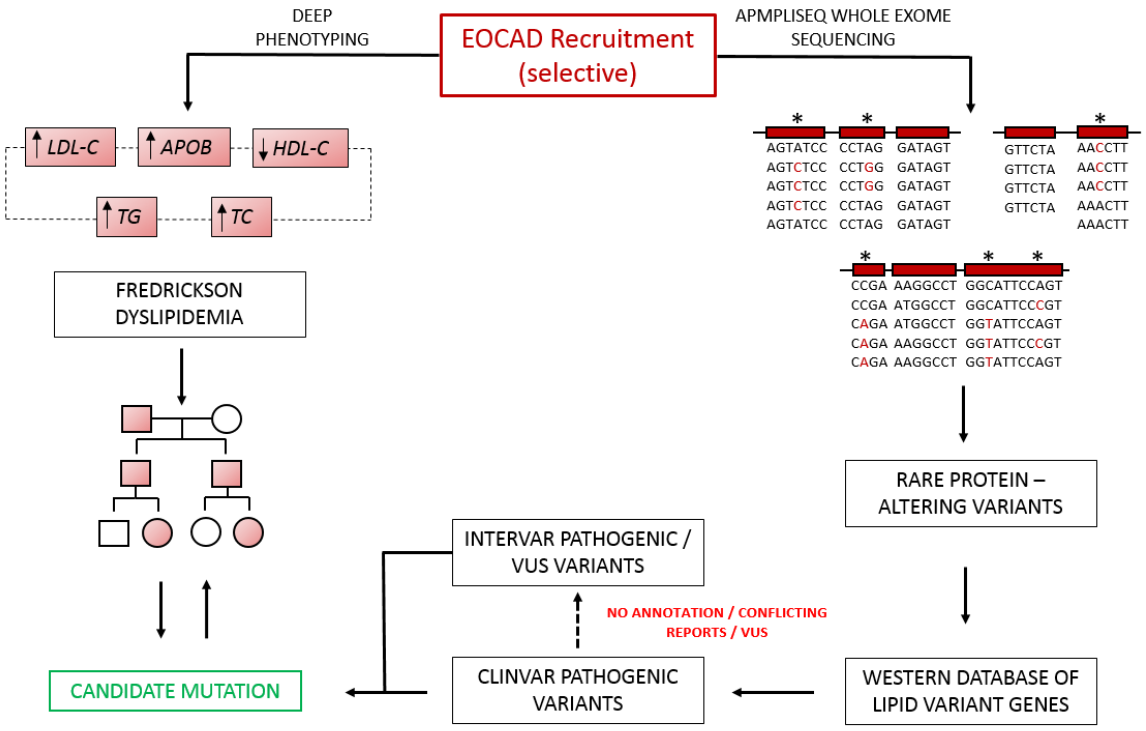


Figure 4.1: Schematic diagram outlining the process of ascertaining variant pathogenicity and matching with carrier phenotype.

4.3 Primer on Familial Hypercholesterolemia (FH)

CAD is a chronic disease caused by the interplay of genetic and environmental factors, and whose heritability is estimated at 40-60% [4,12,67,68]. Numerous genome-wide association studies have been conducted to determine genetic variations that contribute to the development of CAD, including three large meta-analyses (CARDIoGRAM, C4D, CARDIoGRAM plus C4D). Interestingly, a large proportion of variants found to reach genome-wide significance ($P < 5.0x10^{-8}$) involve genes associated with lipid synthesis and transport (Figure 1.3) [12,67,68], highlighting the importance of inborn diseases of cholesterol metabolism to CAD. One such disorder is Familial Hypercholesterolemia (FH), which

exhibits an autosomal dominant mode of inheritance and presents with significantly high levels of plasma cholesterol, a strong risk factor for CAD[58,69]. Specifically, FH results in increased serum LDL-C concentrations due to the presence of mutations in genes that regulate LDL-C transport and metabolism. These mutations exhibit an autosomal dominant Mendelian inheritance pattern and result in increased levels of serum LDL-C by inhibiting receptor-mediated LDL-C endocytosis (Table 4.2). Due to their high penetrance, each mutation is considered diagnostic of FH. However, the severity of disease is largely dependent upon the disruptiveness of the mutation and whether individuals are heterozygous or homozygous for FH mutations (i.e. gene dosage effects [58]). Furthermore, a new form of FH has recently been described which is caused by accumulation of LDL-increasing common genetic alleles of individual weak effect, so-called polygenic FH [70].

The prevalence of DNA-confirmed heterozygous FH is estimated to be 1/500 to 1/300 in the general population and is reported to be as high as 1/77 in non-founder, early-MI populations [58,69,71]. However, these estimations primarily reflect known genetic variants found only in a subset of FH genes, which inherently limits the sensitivity and comprehensiveness that can be obtained through exome-wide sequencing analysis. Nevertheless, enrichment in FH-causing mutations has recently been confirmed in a large case-control study of early-MI, which showed rare, disruptive LDLR mutations to be present in 0.51% of cases and 0.039% of controls, corresponding to an odds ratio of 13 [2]. Not surprisingly, mean LDL-C levels were significantly elevated in mutation carriers as compared to non-carrier controls (7.11 mmol/L vs. 4.52 mmol/L). While these findings support a strong effect of FH mutations on CAD risk, uncertainty on prevalence of FH

mutations in very early CAD populations persists because **(1)** DNA testing was either incomplete or absent, **(2)** many studies also included older participants and **(3)** some studies were done in founder populations. Our study will bridge this knowledge gap by selecting for cases with early onset (median age 38) severe disease, with greater than two-thirds of participants presenting with angiographically-proven multi-vessel disease in our pilot study. Furthermore, we will systematically screen all FH genes for rare mutations and will further assess the presence of polygenic FH. Therefore, our study will provide invaluable insight in this area and will help establish informed clinical criteria for FH screening in this particularly vulnerable patient population. Indeed, the Simon-Broome and Dutch Lipid Clinic Criteria for clinical FH diagnoses are widely used by physicians as they are easy to implement [72]. However, these criteria are unable to definitively diagnose FH in young patient populations in which the classic phenotypes are not yet expressed and can be challenging to apply in secondary prevention patients on lipid lowering therapy [72]. Identification of young FH patients will enable "cascade screening" in affected families.

Table 4.2: Details on known FH genes

Gene Name	Full Name	Function	Most common FH-causing mutation(s)	Mode of Inheritance	Frequency of FH mutations within early CAD/MI	Frequency of rare, damaging & disruptive ExAC *	Ref.
LDLR	Low-density lipoprotein receptor	Binds with APOB ligand on LDL to initiate receptor-mediated endocytosis	Many (~1600)	Autosomal Dominant	Unknown	0.0047	58,69,73
APOB	Apolipoprotein B	Ligand on LDL-C that binds with LDLR to initiate receptor-mediated endocytosis	p.Arg3500Glu	Autosomal Dominant	Unknown	0.0011	58,69
PCSK9	Proprotein convertase subtilisin/kexin type 9	Degradation of endocytosed LDLR	p.Asp374Tyr	Autosomal Dominant	Unknown	5.9×10^{-5} †	58,69
LDLRAP1	Low-density lipoprotein receptor adapter protein 1	Promotes interaction with clathrin-coated pit machinery to facilitate receptor-mediated endocytosis	~10 loss of function	Autosomal Recessive	Unknown	0.0018	58,69
STAP1	Signal transducing adaptor family member 1	Unknown	4 missense	Autosomal Dominant	Unknown	0.0014	69,74
Polygenic FH	Not applicable	LDL-C – raising SNPs	Not applicable	Polygenic	Unknown	0.0026 Φ	70,75

* ExAC consists of over 60 000 exomes from large-scale sequencing studies and can be used to measure **expected** frequencies of the burden of rare variants across genes. The criteria for rare damaging & disruptive variants was: all nonsynonymous variants predicted to be "deleterious" and "probably damaging" by SIFT and Polyphen2, respectively with minor allele frequency < 0.01 and all loss of function (splicing, nonsense, frameshift indel) variants with minor allele frequency < 0.01.

† Exclusive to gain-of-function variants

$\Phi(Prevalenece\ FH\ in\ general\ poulation)\ x\ (Prevalence\ of\ mutation\text{-}negative\ FH)\ x\ (Prevalence\ of\ polygenic\ FH\ in\ mutation\text{-}negative\ FH)$

$\left(\frac{1}{200}\right) x \left(\frac{60}{100}\right) x \left(\frac{88}{100}\right)$

4.4 Short primer on familial combined hyperlipidemia

FCH is often reported as the most common hereditary dyslipidemia with a worldwide prevalence of 1/100 to 1/50 in the general population and as high as 1/5 to 1/10 in early CAD populations [32,76]. However, due to the lack of a consensus clinical or genetic definition of FCH, these are most certainly overestimates. Most typically, however, FCH present with chylomicronemia, hypertriglyceridemia, hypoalphaproteinemia and hypercholesterolemia [32]. Loss-of-function variants within the lipoprotein-lipase (LPL) gene are most commonly associated with FCH and also confer risk for the development of early CAD [24,32].

4.5 Results

4.5.1 WDLV gene coverage

Overall, 92% of WDLV genes (22/24) under investigation had mean % 20X coverage of ≥ 80% (Supplementary Table 2). Only 2 genes (GPIHBP1 and APOE) had sub-optimal % 20X coverage (60.4% and 52.9%, respectively). Mean % 20X coverage for the remaining 22 genes was 89.5 +/- 12.0 %.

4.5.2 Clinical evaluation of FH-mutation positive carriers

Among 53 exome-sequenced DECODE participants, 3 (DECODE 0036, 59, and 68) were found to carry an FH-causing mutation (DECODE FH prevalence = 5.7% (95% CI 1.3-16.0) within the LDLR gene (Table 4.3). All 3 FH-mutation positive cases were males and presented with multi-vessel CAD resulting in severe ACS (2 NSTEMI and 1

STEMI). Additionally, 2/3 FH-mutation positive cases developed CAD very early (DECODE 0036 = 29; DECODE 68 = 35) relative to the FH-mutation negative males in DECODE (median age 38). DECODE 0036 also has history of multiple MIs (1 STEMI + 2 NSTEMI). Significant first-degree family history was also reported in DECODE 59 & 68 whereas DECODE 0036 only reported an early MI event in a maternal aunt. All FH-mutation positive cases are smokers. DECODE 68 also presents with hypertension (127 mmHg/76 mmHg while on anti-hypertensives) and DECODE 0036 was diagnosed with type II diabetes mellitus 2 years prior to his first event and is currently undergoing non-insulin-dependent treatment. All FH-mutation positive cases presented with type IIa hyperlipidemia with LDL-C values of 4.1, 4.4 and 3.8 mmol/L for DECODE 0036, 59, and 68, respectively (4.1 +/- 0.39 mmol/L) after correcting for statin treatment based on statin type and dosage regimen according to well-established adjustment coefficients (Table 4.3).

Table 4.3: Statin-adjusted LDL-C for FH-mutation positive carriers in DECODE

DECODE ID	LDL-C (mmol/L)	Statin	Adjustment coefficient	Statin-adjusted LDL-C (mmol/L)
0036 †	3.46	80mg Simvastatin	1.46	5.05
59 Ψ	4.40	None	-	4.40
68	2.68	40 mg Rosuvastatin	1.55	4.15

† Patient also on 10 mg Ezetrol; Ψ Patient not on statin at time of blood draw

DECODE 0036 harbours a nonsense variant within exon 14 (rsID = rs121908031; c.2043C>A; p.Cys681X) (Table 4.4) which encodes a portion of the epidermal growth factor (EGF) precursor homology domain in the LDLR protein. The variant in question results in a truncated LDLR characterized by absence of both the membrane-spanning and

cytoplasmic domains. This variant represents a founder mutation known as the Lebanese allele (FH Lebanese) as it exhibits higher frequency amongst Lebanese Arabs, consistent with this DECODE participant's ethnicity. This variant is annotated as "pathogenic" within ClinVar based on 8 independent submissions (Table 4.3). DECODE 59 and 68 carry missense variants (rs551747280; c.82G>T; p.Glu28Lys & rs397509365; c.1690A>C; p.Asn564His) (Table 4.4), within exon 2 and 11 of LDLR, respectively. Exon 2 encodes a portion of the LDL receptor class A domain repeats which are responsible for interaction of with the Apo-B ligand expressed on plasma LDL and exon 11 encodes a subset of the EGF precursor homology domain responsible for acid-dependent dissociation of LDL-C from the internalized LDL-LDLR complex. The rs551747280 variant observed DECODE 59 was only recently annotated contain a ClinVar annotation and was therefore manually queried using InterVar to achieve standardized pathogenicity interpretation (Table 4.4). The variant was found to be positive for 2 ACMG/AMP guidelines that support pathogenicity and 1 that supports it being benign: PP2, PP4, and BS2 (described in Supplementary Table 3) which resulted in a VUS interpretation. However, given the extensive family history of this individual along with having a concordant phenotype for FH (i.e. elevated LDL-C and early CAD), it is likely that this variant is causal. The rs397509365 variant in DECODE 68 variant was reported as both "likely pathogenic" and "pathogenic" in ClinVar according to 4 independent submissions (Table 4.4). Similar to rs551747280 (Lebanese allele), rs397509365 is also a founder mutation known as the Aarhus allele (FH Aarhus) and exhibits higher frequency in Dutch Europeans. The Aarhus allele was also accompanied with a 9 base pair (nonframeshift) deletion (in *cis* to the

missense variant) located in exon 17 (c.2393_2402delTCCTCGTCT; p.Leu799_Phe801del). Evidence from in vitro analysis has demonstrated that the presence of both the missense and deletion variants are necessary to reduce LDL-C uptake by 75% in transfected COS cells, suggesting a synergistic effect.

Both rs551747280 and rs397509365 were found to be "pathogenic" according to their M-CAP score. Causal variants within the LDLR gene are segregated into five classes which define their functional effect. Class descriptions for LDLR variants discovered in DECODE are summarized in Table 4.5.

Table 4.4: Summary of variants predicted to be causal for Mendelian dyslipidemias in the DECODE cohort.

ID	Ethnicity	Interim diagnosis	Gene	Variant ID	ExAC global frequency	Mutation type	ClinVar annotation	Positive InterVar criteria Φ	Final status
0036	Arab	Familial hypercholesterolemia *FH Lebanese*	LDLR	rs121908031	8.255e-06	Nonsense	Pathogenic - FH	-	Pathogenic
59	European	Familial hypercholesterolemia	LDLR	rs551747280	0.0002891	Missense	No annotation	PP2, PP4, BS2	VUS
68	European	Familial hypercholesterolemia *FH Aarhus*	LDLR	rs397509365	0	Missense	Pathogenic - FH	-	Pathogenic
				NA	0	Non-frameshift	NA		
20	European	Familial combined hyperlipidemia	LPL	rs268	0.01336	Missense	Pathogenic - FCH	-	Pathogenic
42	European	Familial combined hyperlipidemia	LPL	rs268	0.01336	Missense	Pathogenic - FCH	-	Pathogenic
60	European	Familial combined hyperlipidemia	LPL	rs1801177	0.01492	Missense	Pathogenic-FCH	-	Pathogenic

Φ Description of criteria codes for InterVar used to standardize pathogenicity ranking are described in Table S3.

Table 4.5: Summary of the LDLR mutation classes.

Variant ID	LDLR mutation class	Class Description
rs121908031	**4**	Defect in LDLR clustering on cellular membrane
rs551747280	**3**	Diminished affinity between LDLR and APOB ligand on plasma LDL-C
rs397509365	**5**	Defect in acid-dependent dissociation of the internalized LDLR-LDL-C complex which compromises LDLR recycling to cellular membrane

4.5.3 Enrichment of FH in the DECODE cohort relative to CAD-free controls and an unselected patient population

When defining FH as having both hypercholesterolemia (LDL-C > 3.36 mmol/L) and a rare (MAF < 0.01) missense (predicted pathogenic by *in silico* pathogenicity tools or by ClinVar) or rare disruptive LDLR variant, the presence of an FH mutation was associated with a 30-fold (95% CI 3-128) increase risk of early CAD when compared to a reference population consisting of CAD-free and unselected participants in the MI-Gene + CHARGE consortia (P=0.002) (Table 4.6). When restricting analysis to only CAD-free controls and instituting a more widely accepted genetic definition for FH (rare disruptive LDLR variant + rare missense LDLR variant predicted to be pathogenic by ClinVar) while withholding threshold for LDL-C, FH mutation carriers were associated with a 20-fold (95% CI 2-87) increased risk of early CAD (Table 4.6). The effect observed here is approximately 3-fold greater than the effect observed by Abul-Husn et al. 2016 [77] on early CAD (Figure 4.2). No significant enrichment in FH mutation carriers was identified for early CAD when defining FH solely on the basis of severe hypercholesterolemia (i.e. LDL-C > 4.9 mmol/L) (P>0.05).

Table 4.6: Association of FH mutations in DECODE compared to the MIGen + CHARGE consortia

	DECODE	MIGen + CHARGE consortia	OR (95% CI) FH+ vs FH-	P-value FH+ vs FH-
Statin corrected LDL-C ≥ 3.36 mmol				
N (FH mutation pos †)	2	68	30.0 (3.4 – 127.8)	0.002
N (FH mutation neg)	20	20417		

	DECODE	MIGen CAD-free controls	OR (95% CI) FH+ vs FH-	P-value FH+ vs FH-
No LDL-C threshold				
N (FH mutation pos Ψ)	2	17	19.7 (2.2 - 86.7)	0.005
N (FH mutation neg)	51	8561		

† Rare disruptive LDLR variants + rare missense LDLR predicted pathogenic by each of 5 *in silico* pathogencity criteria (SIFT, Polyphen2 HumDiv, Polyphen2 HumVar, MutationTaster, LRT) + rare missense LDLR variants predicted to be pathogenic in ClinVar; Ψ Rare disruptive LDLR variants + rare missense LDLR variants predicted to be pathogenic in ClinVar

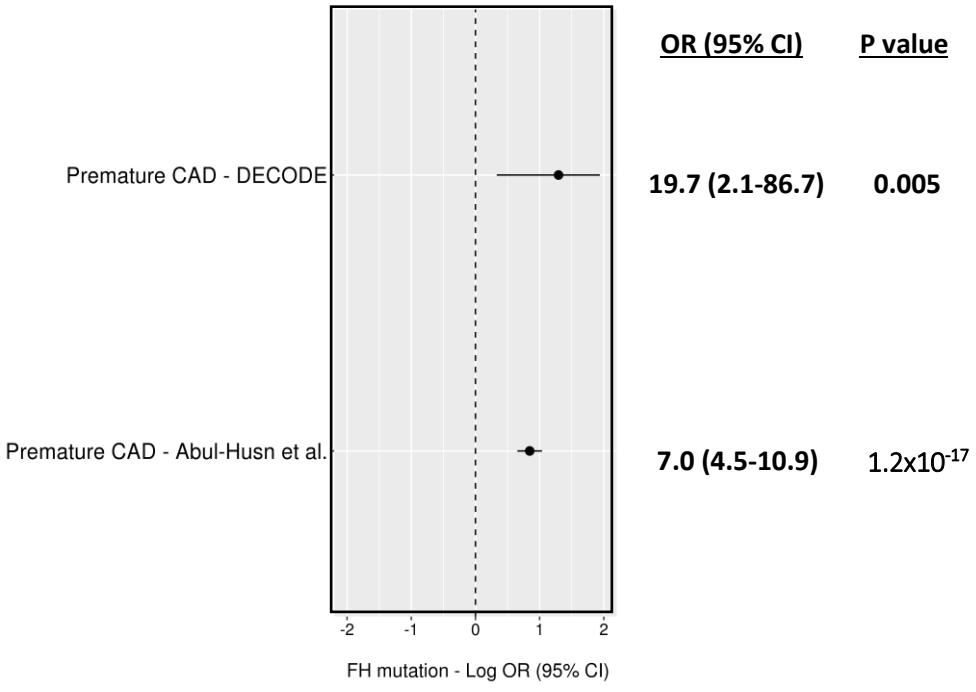


Figure 4.2: Association of premature CAD with FH mutations in analysis conducted by Abul-Husn et al. and DECODE. CAD was defined as premature in males < 55 and females < 65 for the Abul-Husn et al. analysis. CAD was defined as premature in males < 40 and females < 45 for in DECODE. ORs were calculated using logistic regression with adjustment for age, sex and ethnicity in the Abul-Husn analysis and using 2-tailed fisher exact test in DECODE. FH mutations were defined as disruptive LDLR variants + missense LDLR variants predicated to be pathogenic in ClinVar.

4.5.4 Clinical evaluation of FCH-mutation positive carriers

Two DECODE participants (20, 42) are carriers of a single missense mutations within LPL (rs268; c.953A>G; p.N318S) (Table 4.4) which is annotated as pathogenic in ClinVar and causal for FCH. rs268 has a global ExAC frequency of > 0.01 and consequently does not have a M-CAP annotation. However, the CADD score for this variant is 21 which corresponds to the 99th percentile of deleterious variants. DECODE 20 and 42 both have elevated LDL-C levels at 4.81 and 3.59 mmol/L, respectively. DECODE 42 was also on statin treatment (80 mg atorvastatin) at time of blood draw. Both participants also display modestly elevated triglyceride levels at 2.01 and 2.22 mmol/L and decreased HDL-C levels at 0.75 and 0.78 mmol/L, respectively. Collectively, the lipid panel for both individuals is most consistent with type IIb dyslipidemia. The rs268 variant has recently been shown to confer the highest effect for increased triglycerides amongst low frequency and common variants (beta = 0.17 mmo/L per risk allele) and is also independently associated with CAD.

An additional LPL variant was found in DECODE 60 (rs1801177, c.9126G>A, p.D36N) and has previously been reported as pathogenic for FCH in ClinVar, however the lipid panel for this individual is incomplete with no additional records on our electronic

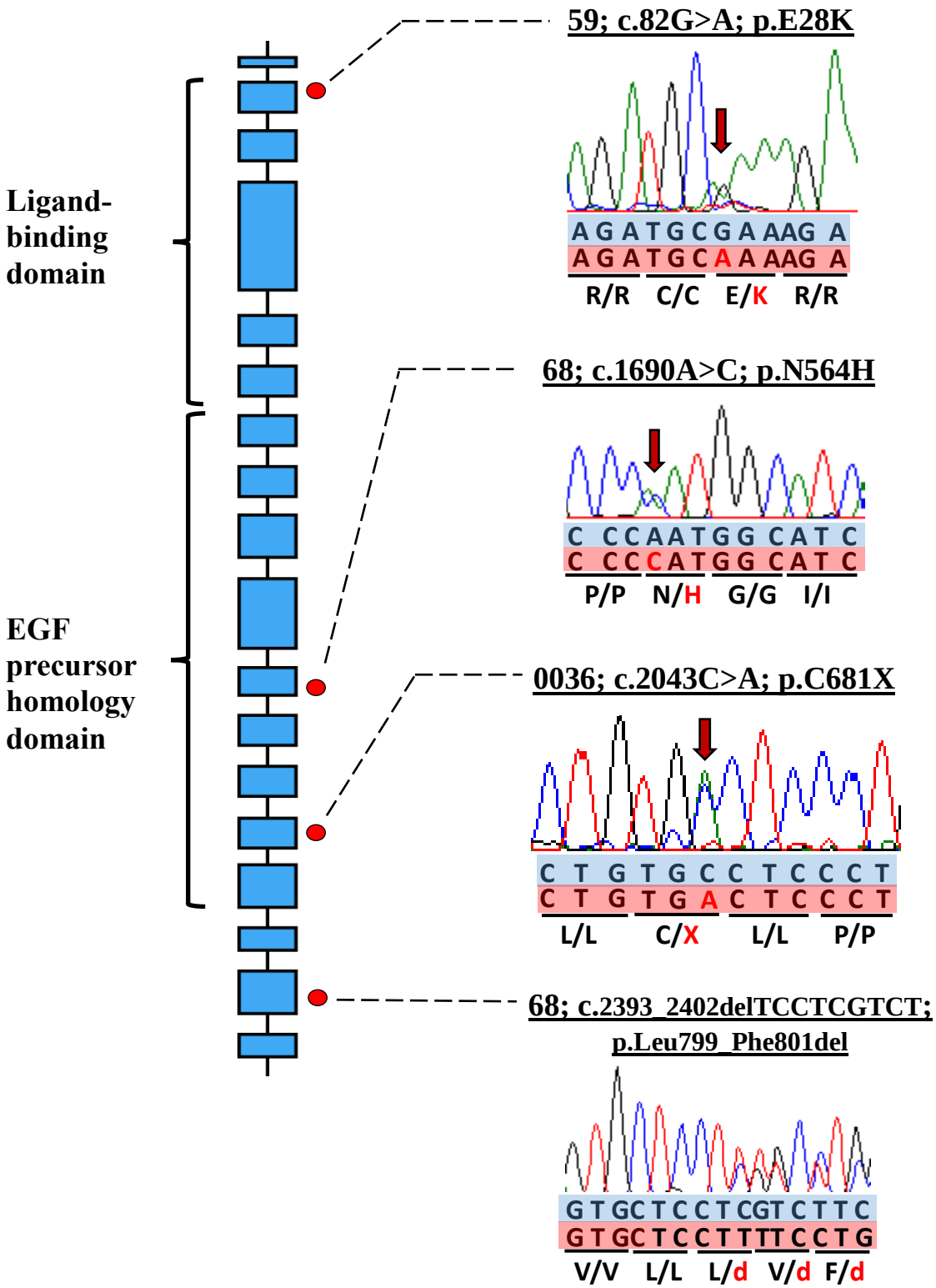
health databases. We do declare this variant in our summary results (Table 4.4), but fully acknowledge that it is difficult to ascertain whether it can be classified as causal.

4.5.5 Diagnostic yield

When taking into account all FH and FCH variants predicted to be disease-causing in our pipeline, we ascertain a diagnostic yield of 10%, with all cases being clinically actionable.

4.5.6 Sanger sequencing validation

All variants discussed were confirmed using Sanger sequencing according to workflow described in section 4.2.5. Figure 4.3 depicts schematics for the LDLR and LPL genes annotated with the variants discussed. Electropherograms confirming presence of variants in patient samples are also provided.


Ligand-binding domain
EGF precursor homology domain
59; c.82G>A; p.E28K
A G A T G C G A AAG A
A G A T G C A A AAG A
R/R C/C E/K R/R
68; c.1690A>C; p.N564H
C CC AA T G G C A T C
C CC C A T G G C A T C
P/P N/H G/G I/I
0036; c.2043C>A; p.C681X
C T G T G C C T C C C T
C T G T G A C T C C C T
L/L C/X L/L P/P
68; c.2393_2402delTCCTCGTCT; p.Leu799_Phe801del
G T GC T C C T CGT C T T C
G T GC T C C T T TT C C T G
V/V L/L L/d V/d F/d

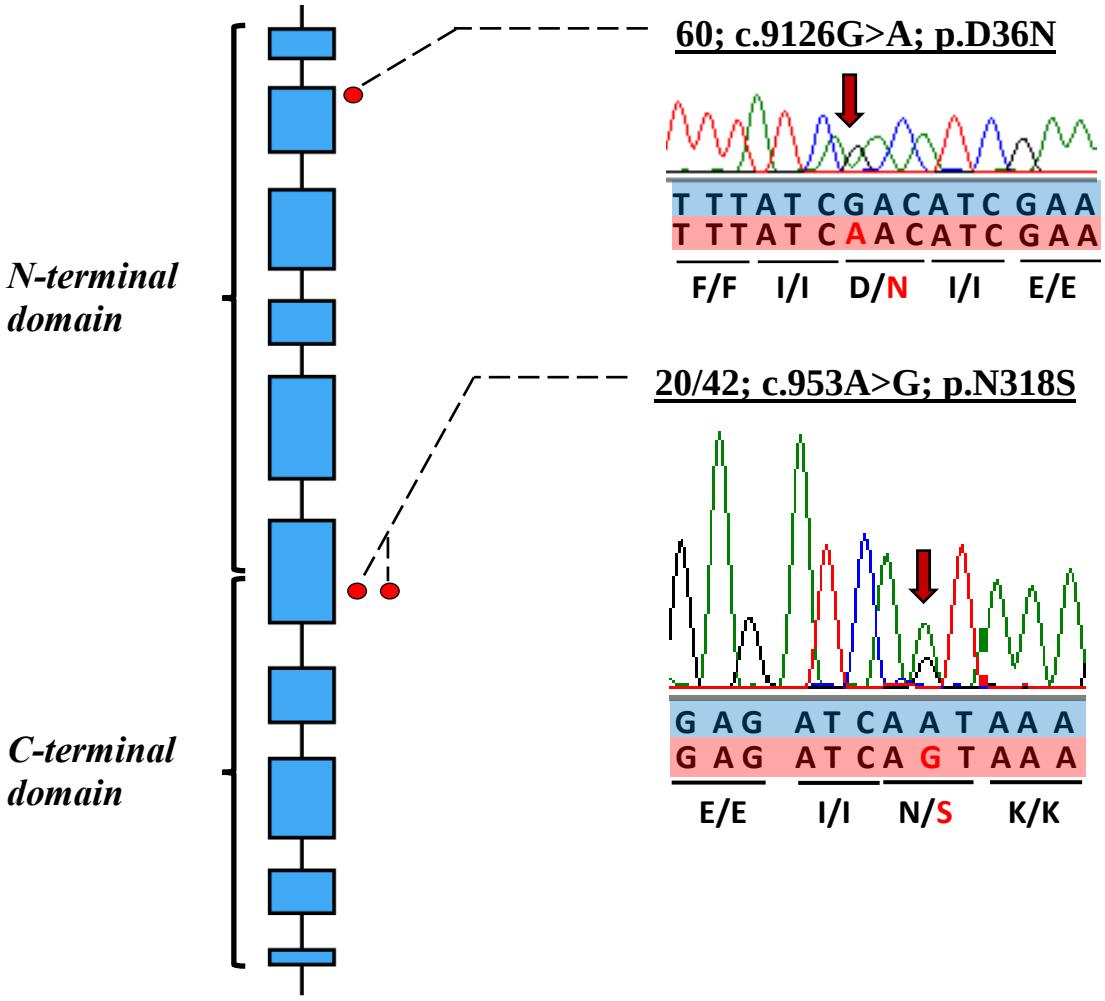


Figure 4.3: Schematics of the LDLR (A) and LPL (B) gene along with the electropherograms depicting the variants (red arrows) causal for FH and FCH, respectively. The header for the electropherogram describes the sample IDs for mutation carriers, coding sequence variation and protein variation, respectively. Nucleotides highlighted in blue and red represent the reference and alternate sequences, respectively. Black bars encompass individual codons with the amino acids encoded by reference and alternate sequence provided below (separated by a slash).

4.6 Discussion

FH and FCH represent the two most common inherited disorders in the general population and are significant risk factors for the development of premature CAD [32,61,76,77]. Both conditions are critically under-diagnosed (especially in the Western world) based on expected prevalence estimates established by countries with health policies in place to conduct systematic genetic screening. For example, using the conservative global prevalence of FH (1/500), it is expected that ~600000 Canadians currently present with sufficient clinical and/or genetic evidence to support FH diagnosis. However, less than 1% of these cases have currently been diagnosed which provides strong rationale to institute genetic screening early in life in order to facilitate timely pharmacological or lifestyle interventions and initiate cascade screening in families with individuals harbouring FH-causing variants.

The issue of unknown FH prevalence is particularly concerning among individuals with early CAD, who by virtue of their stringent cholesterol-lowering treatment regimens, do not meet clinical requirements (i.e. highly elevated LDL-C and total cholesterol) necessary to warrant a consensus FH diagnosis. However, it has been shown that FH-mutation positive carriers demonstrate substantially higher risk for CAD compared to FH-mutation negative individuals across all strata of LDL-C levels (e.g. FH-mutation positive individuals with LDL-C < 3.3 mmol/L were at a 2-fold higher risk for developing CAD compared to individuals who are FH-mutation negative within the same LDL-C strata) [61]. This observation underscores the necessity of incorporating genetic information to appropriately assess disease risk. The increased in CAD risk for FH-mutation positive

individuals can largely be attributed to cumulative, life-long exposure to elevated LDL-C as opposed to acute elevations likely related to diet and physical activity which are not sufficient to fully develop atherosclerotic plaques that are characteristic of CAD. This finding is strengthened when considering that all FH-mutation positive carriers in DECODE failed to reach LDL-C levels recommended by the National Lipid Association's Expert Panel [78] for adults with FH and CAD (< 1.8 mmol/L) even while on statin treatment. Although this observation is based on a small sample size, evidence in the literature from cohorts encompassing thousands of individuals support the discrepancy between FH-mutation positive and negative individuals in the ability to reach desirable LDL-C while on statins [77].

In order to validate our 3rd hypothesis in this work, we set to evaluate the effect and putative association between CAD status and frequency of an FH mutation. We find that FH mutations were associated with a 30-fold increased for early CAD accompanied with hypercholesterolemia (LDL-C > 3.36 mmol/L) when using CAD-free controls and unselected participants from prospective population cohorts as the reference (both also with LDL-C > 3.6 mmol/L). This effect illustrates that individuals with an FH mutation hold substantially greater risk for developing early CAD as compared to a reference population that is similarly hypercholesteremic. This is consistent with the observation (as stated earlier) purported by Khera *et al.* 2016 [61] in which individuals with similar LDL-C were evidently more likely to develop CAD if they were FH-mutation positive.

We next sought to compare the difference in FH mutation frequency between DECODE and other early CAD cohorts using only CAD-free individuals in MIGen as

reference. A recent analysis published by Abul-Husn *et al.* 2016 [77] observed that FH mutations (see Table 4.3 for pathogenicity criteria) were associated with a 7-fold increased risk of early CAD (defined as males ≤ 55 and females ≤ 65) compared to only a 4-fold increased risk seen in general CAD. In DECODE, however, FH-mutations were associated with a 20-fold increase in early CAD risk, which is quite similar to the effect observed in a separate study between FH-mutation positive carriers with severe hypocholesteremia (LDL-C > 4.9 mmol/L) and FH-mutation negative individuals that were normocholesterolemic (LDL-C < 3.3 mmol/L). The discrepancy between effects observed in young populations may largely be attributed to the upper age limits used to define "early" CAD. In the aforementioned study, males ≤ 55 and females ≤ 65 were considered to have early disease whereas the median age for males and females in DECODE is 37 and 41 respectively. We therefore establish that frequency of FH mutations increases extensively in a manner that is largely age-dependent, which can serve as a proxy for disease severity in the context of CAD.

Collectively, these observations point to the urgent need to perform systematic genetic-based FH screening in early CAD cohorts in order to appropriately modulate lifestyle choices and optimize pharmacological interventions. Conducting FH screening on the basis of clinical criteria alone is in insufficient and uncomprehensive, especially when considering that < 2-2.5 % of individuals carrying an FH-causing mutation have severe hypercholesterolemia as reported by two recent large clinical analyses [61,77]. It was also shown that only 1/4 individuals with an FH-mutation were diagnosed with possible or definite FH when using the Dutch Lipid Network criteria on the basis of clinical electronic

health record data alone [77]. The extremity of these findings are further amplified in early CAD cohorts as we show significant increases in FH-mutation frequency to be representative of these populations.

The individuals within DECODE found to carry FH mutations have been contacted to follow-up appointments in our Lipid Clinic to re-assess type and dosage of lipid-loweirng medication. Additionally, we have received permission to perform "cascade testing" in family members to evaluate potential genotype-phenotype co-segregation. This can inform appropriate clinical and lifestyle interventions in close relatives which may possibly prevent severe clinical outcomes.

We also acknowledge that the confidence intervals observed in our results are a manifestation of the small sample size of the DECODE study. We fully expect this to be ameliorated over the next several months as more individuals become recruited. Nevertheless, we observe that the direction and magnitude of effect to be as expected.

CHAPTER 5 – Rare Variant Association: Leveraging External and Internal Control Datasets for Gene Discovery

5.1 Introduction

Rare variant association analyses (RVAS) have proven invaluable for the detection of genes causally linked to CVD that would not have otherwise been identified through standard GWAS [2,25,26,30]. The success of RVAS is grounded in its study design. Methodologies including EPS and biological validation of discovered variants (both of which are employed in this work) can vastly empower studies to detect associations of large effect with smaller sample sizes. Nevertheless, the advent of large external, publicly available sequence databases (e.g. ExAC and gnomAD) have made it possible to substantially increase the sample size and power of association analyses without incurring large sequencing costs. Due to their sheer size, external sequencing datasets can be used as control populations to facilitate identification of rare variants within the case population (i.e. case-only analyses). This is especially attractive for studies examining rare or early phenotypes which are inherently limited in their ability to reach large sample sizes. However, studies that employ the use of external databases as control sets require rigorous adjustments to account for potential differences in sequencing chemistries, which may lead to spurious association signals due to technical artifacts. Moreover, external sequencing databases can be phenotypically heterogeneous and it must be ensured that these datasets are large enough to dilute variants that could otherwise mask true association signals. External sequencing databases may also provide only summary-level data on called variants which requires use of alternate statistical methodologies to assess the impact of rare variant burden on disease risk.

The utility of case-only analyses for the discovery of genes causally linked to complex diseases such as CAD is currently unknown. We herein leverage the large sample size of the ExAC database (N=60706) as our reference control set to identify significant or nominal gene-based associations with early CAD. We also use internally sequenced, CVD-free controls from the ORIGIN trial [79] (N=409) as a means of controlling for potential residual biases. Using internal controls also offers the advantage of having access to individual-level data, which may allow for **1)** stringent selection of samples that are not afflicted with CAD and access and **2)** employ the use of variance component tests (e.g. SKAT) that may result in discovery of novel genes that could not be detected with traditional rare variant burden testing.

5.2 Methods

5.2.1 DECODE study population

The DECODE population is described in detail in section 2.3.1.

5.2.2 Early-Onset Myocardial Infarction (EOMI) study population

A total of 736 early-MI cases were obtained from The Early-Onset Myocardial Infarction (EOMI) cohort within the NHLBI GO ESP6500 via application to the database of Genotypes and Phenotypes (dbGAP) (study accession: phs000279.v2.p1) [80]. EOMI cases were drawn from 5 community based studies (PennCATH, Cleveland Clinic Genebank, Massachusetts General Hospital Premature Coronary Artery Disease Study (MGH-PCAD), Heart Attack Risk in Puget Sound (HARPS), and Translational Research

Investigating Underlying Disparities in Myocardial Infarction Patients' Health Status (TRIUMPH)). The proportion of cases recruited from each study is provided in Supplementary Table 4. The case definition for EOMI included an MI (STEMI or NSTEMI) in males ≤ 50 and females ≤ 60.

5.2.3 ExAC

The ExAC dataset consists of 60706 exomes which were aggregated across 17 exome sequencing datasets (Supplementary Table 4) [20,21]. Version 0.3 of ExAC dataset (ExAC v0.3) was used as the reference population for case-only association analyses. The VCF containing 10.2 million variant calls (92% SNV; 8% INDEL) along with coverage files were downloaded from the ExAC ftp repository: (ftp://ftp.broadinstitute.org/pub/ExAC_release/release0.3/). The raw VCF was filtered to retain only sites receiving a PASS filtering annotation (retaining 9.1 million variant sites - 94% SNV, 6% INDEL) prior to any downstream analyses.

5.2.4 The ORIGIN trial

The ORIGIN trial has been described previously [79]. Briefly, 12,537 participants with cardiovascular risk factors and evidence of dysglycemia were randomized in a 2x2 factorial design to either oral insulin glargine versus standard care, and omega-3 fatty acid supplementation versus placebo, and followed for a median of 6.2 years. A total of 494 ORIGIN participants have undergone exome sequencing in our laboratory. We have selected 409 of these individuals who have not demonstrated history of CVD as a control population for SKAT association testing.

5.2.5 Sample-level QC for the EOMI study population

Kinship analysis was conducted exclusively on common (MAF > 0.01) SNVs that were LD-pruned according to a window size of 100, window shift of 50 and an r^2 threshold of 0.2 using plink v1.9. Pairwise kinship coefficients (k_0) between all samples was calculated using KING [81]. Duplicate samples ($k_0 > 0.354$) or sample pairs demonstrating first ($0.177 < k_0 < 0.354$), second ($0.177 < k_0 < 0.0884$), or third ($0.0442 < k_0 < 0.0884$) degree relatedness were flagged for removal. The sample demonstrating the higher overall call rate from each pair was retained. Sex check and heterozygosity checks were also performed using plink v1.9. Samples with discordance between reported and genetically determined sex as well as samples demonstrating excess heterozygosity were removed.

5.2.6 Variant-level QC for the EOMI study population

All variants were assessed for Hardy-Weinberg Equilibrium (HWE) using an exact test as defined by Wigginton *et al.* 2015 [82]. This test was implemented in vcftools [83]. Variants demonstrating nominal deviation from Hardy-Weinberg Equilibrium ($P<0.05$) were removed from downstream analysis.

5.2.7 Exome library preparation and exome sequencing

Exome library preparation and exome sequencing were performed as described in sections 2.4.3-5.

5.2.8 Variant calling and annotation

Variant calling and annotation was performed as described in sections 2.4.7-8.

5.2.9 Variant pathogenicity filtering

Called variants passing quality thresholds in were annotated with their corresponding MAFs observed in 1KGP3 [50], NHLBI GO ESP6500 [51] , and ExAC [20] for rare variant filtering. Variants annotated with a MAF < 1% (T1 alleles) and MAF < 5% (T5 alleles) in all ethnicities among all external exome databases (1KGP3, NHLBI GO ESP6500, ExAC v0.3) were kept. T1 and T5 alleles were further filtered for MAF < 0.01 and MAF < 0.05, respectively within an in-house dataset consisting of 248 samples sequenced using the Ion Proton™ and S5XL™ systems in order to account rare variant sequencing artifacts. Rare variants subsequently underwent functional annotation using several *in silico* protein-prediction algorithms (using dbNSFP v.3.0 [48]) in order to effectively discriminate neutral from putative disease-causing variation according to four pathogenicity schema: 1) all variants resulting in mutation types that alter the primary protein sequence (i.e. protein-altering) 2) All disruptive variants (splicing, stopgain, stoploss, frameshift indel) in addition to all nonsynonymous SNVs predicted to be "deleterious" and "damaging" according to SIFT and Polyphen2-HDIV/HVAR, respectively 3) All disruptive variants in addition to nonsynonymous SNVs with a CADD score > 20 4) All disruptive variants in addition to all nonsynonymous SNVs predicted to have an M-CAP score > 0.025. Definitions for all mutation and pathogenicity types can be found in Tables 2.3 and 2.4, respectively. Identical frequency and functional filtering were applied to the ExAC reference dataset in order to associate the aggregate frequency of rare, disease-causing variants among all genes with the early-onset CAD phenotype.

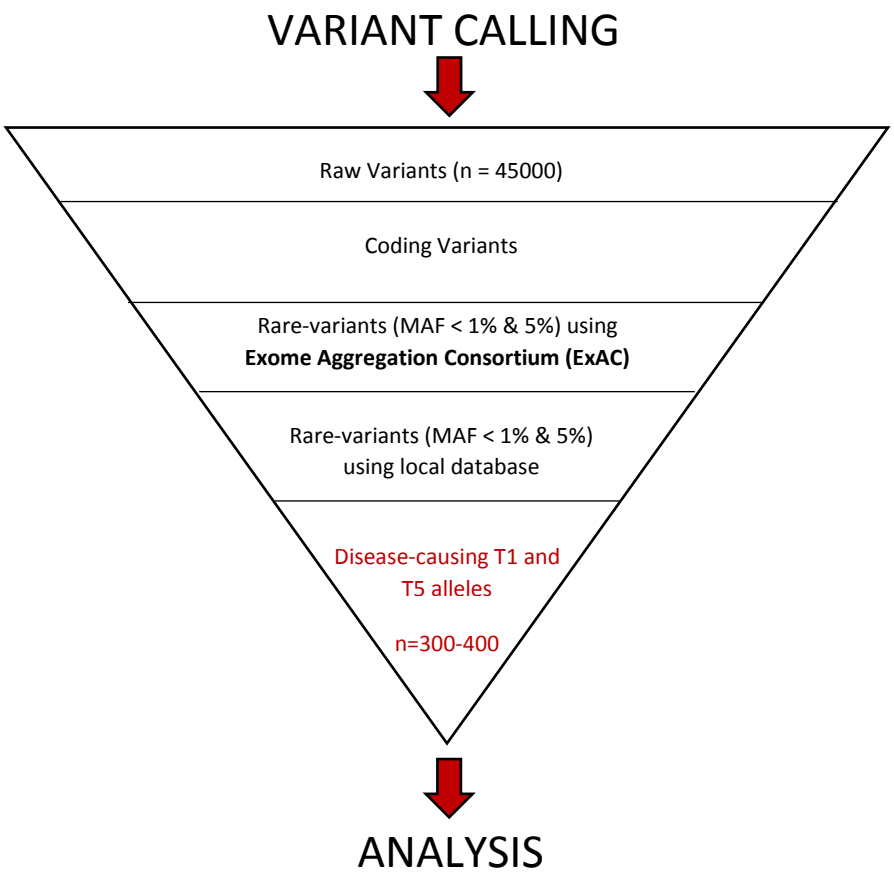


Figure 5.1: Flow-diagram outlining filtering criteria necessary to achieve putative set of rare, disease-causing variants that can be used in association analyses.

5.2.10 Association models

Gene burden tests for T1 and T5 alleles were conducted using additive, dominant, and recessive models in order to determine if specific genes were conferring disease risk through different modes of inheritance (Figure 5.2).

In the additive and dominant models, aggregate allele (additive) and carrier frequencies (dominant) were determined for T1 and T5 alleles across all stringencies in DECODE and ExAC cohorts. The dominant model functions as a sensitivity analysis to the additive

model as it accounts for association signals that may be driven by only one or two carriers. For the recessive model, aggregate **recessive** carrier frequency were determined for homozygous and *trans* compound heterozygous T5 alleles across all stringencies in DECODE. *Trans* compound heterozygous variants were defined as distinct heterozygous variants that lie on separate homologous chromosomes. These were identified through phasing genotypes to their respective homologous chromosomes using the hidden markov model-based SHAPEIT2 algorithm [84]. In order to facilitate phasing accuracy, both the 1KGP3 and Haplotype Reference Consortium (HRC) [85] were used as reference panels. 1KGP3 and HRC consist of 82 million and 40 million variant sites across 5008 and 64940 haplotypes, respectively and are therefore ideal for robust phasing of both rare and low frequency variants. In ExAC, recessive carrier frequencies were calculated based on the square of the carrier frequencies determined in the dominant analysis as this provides the expected frequency of observing an allele twice in a given individual.

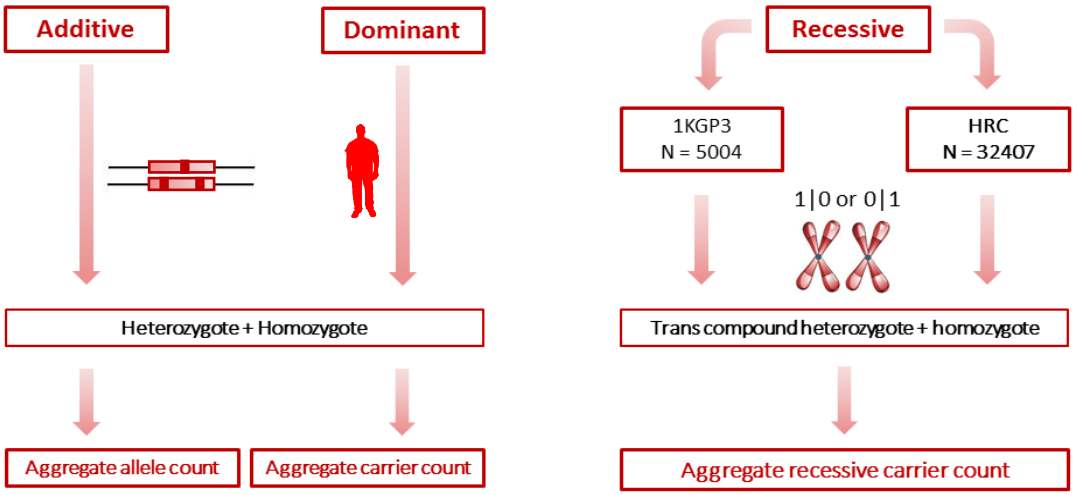


Figure 5.2: Schematic outlining the additive, dominant, and recessive association models.

5.2.11 Coverage adjustments for external controls

Briefly, mean % 20X coverage was determined for each exon in both DECODE and ExAC. Exons exhibiting differential % 20X coverage of > 10% were removed from analysis. This is described in further detail (for single samples) in section 6.2.4.

5.2.12 Statistical analyses using external controls (ExAC)

Gene burden tests were used to evaluate the *in-aggregate* frequency (i.e. cumulative minor allele frequency (CMAF)) of T1 and T5 alleles across all pathogenicity rankings (see section 5.2.7) for every individual gene assessed within DECODE, EOMI and ExAC cohorts. Gene based p-values were calculated using the cumulative binomial probability distribution function in R [86]. A weighted-minor allele frequency and minor allele carrier frequency ($wMAF$ and $wMACF$, respectively) (equation 5.1 and 5.2) that are based on the ethnic structure of DECODE were determined using the MAF of each T1 and T5 allele within ExAC:

$$wMAF = \sum_{e=1}^{M} \hat{p}\,(DECODE)_e\; ExACMAF_e \qquad \text{Equation 5.1}$$

$$wMACF = \sum_{e=1}^{M} \hat{p}\,(DECODE)_e ExACMACF_e \qquad \text{Equation 5.2}$$

where $\hat{p}\,(DECODE)$ represents the proportion of ethnicity e in the DECODE study and $ExACMAF$ and $ExACMACF$ represent the allele and carrier frequencies, respectively, for

a given variant in ethnicity e in ExAC. The $wMAF$ and $wMACF$ meeting pre-specified pathogenicity criteria for all alleles were subsequently summed for each gene and used as the expected, gene-specific CMAF (additive) and cumulative minor allele carrier frequency (i.e. CMACF) (dominant) obtained by chance. For the recessive association model, we used the square of the CMACF as this provides the expected frequency of observing an allele twice in a given individual. The expected CMAF/CMACF for each gene was subsequently measured against the observed T1 and T5 allele/carrier counts in DECODE using a sample size of 104 alleles for the additive analysis and 52 cases for the dominant and recessive analysis in order to determine the cumulative probability of observing an aggregate frequency higher than what is expected in ExAC (i.e. the right-tailed probability) using the survival function:

$$P\left(DECODE\ aggregate\ allele\ frequency > ExAC\ aggregate\ allele\ frequency\right)_j$$

where j represents each individual gene. Exome wide significance was uniquely determined according the "*discrete method*" of multiple hypothesis correction for discrete distributions described by Westfall and Wolfinger [87] for all association tests. Briefly, discreteness is incorporated into multiplicity adjustments by sampling the maximum permuted p-values p_{per} across m genes (based on N observances; where N = number of alleles or carriers) that fall below a significance threshold p_{thresh} to produce a family-wise error rate (FWER) of 0.05 according to:

$$FWER(0.05) = 1 - \prod_{j=1}^{M} 1 - p_{minimum} \qquad \text{Equation 5.3}$$

where

$$p_{miniumum} = \begin{cases} max(p_{per} < p_{thresh}) & if\ min\ p_{per} < p_{thresh} \\ 0, & otherwise \end{cases}$$

P-value cutoffs for exome-wide significance are provided in Supplementary Table 6 and 7 for each permutation of association model, MAF threshold and pathogenicity requirement. Gene-based p-values < 0.01 were deemed as being nominally significant. All statistical computations for this section were conducted in R version 3.2.2 unless otherwise stated.

5.2.13 Statistical analysis using internal controls

The Sequence Kernal Association Test (SKAT) was used to regress each variant in a gene on the binary phenotypic outcomes of: CAD and CAD-free using the skatMeta package in R [86,88]. Outcomes were adjusted for age, sex, and BMI. Variants exhibiting differential missingness between case and control populations were removed from analysis.

5.3 Results and Discussion

5.3.1 Ethnic composition of the DECODE cohort

The ethnic composition of the DECODE cohort is fully outlined in section 2.5.2.

5.3.2 Ethnic composition of the EOMI cohort

The 736 individuals comprising the EOMI cohort generated by NHLBI GO ESP6500 are consisted of 560 (76.0%) Europeans, 146 (19.8%) Africans, 14 (19.8%) Latin Americans, and 28 (3.8%) members of minor or unknown ethnicities. Due to their sample sizes, only Europeans and Africans were used in association analysis.

5.3.3 Ethnic composition of the ExAC database

The 60706 participants within the ExAC dataset consisted of 33370 (55.0%) Non-Finnish Europeans (NFE), 5203 (8.6%) Africans (AFR), 5789 (9.5%) Latin Americans (AMR), 4327 (7.1%) East Asians (EAS), 3307 (5.4%) Finnish Europeans (FIN), 8256 (13.6%) South Asians, and 454 (0.75%) members from minor ethnic groups.

5.3.4 Additive analysis identifies known CAD GWAS genes in EOMI Europeans using the ExAC NFE population as the control dataset

After interrogation of genes enriched for disease-causing T1 and T5 alleles in the EOMI European cohort, we detected an exome-wide significant associations of Cadherin EGF LAG seven-pass G-type receptor 2 (CELSR2) ($P = 1.1x10^{-17}$) and a nominal association of the Apolipoprotein A-V (APOA5) (P = 0.001) (Table 5.1) with MI.

5.3.4.1 CELSR2

The association signal for CELSR2 was driven by 17 T5 nonsynonymous SNVs across 76 heterozygous and 1 homozygous carrier(s). All variants were predicted to be damaging using the M-CAP *in silico* pathogenicity tool. CELSR2 is a member of the

adhesion G protein-coupled receptors and is expressed ubiquitously, with highest expression shown in the liver and pancreas [89,90]. An intergenic SNP (rs599839) localized to the CELSR2-PSRC1-SORT1 locus has independently associated with CAD/MI and LDL-C in recent GWAS meta-analyses at genome-wide significance ($P = 2.9x10^{-10}$ for CAD/MI; $P = 1.8x10^{-11}$ for LDL-C) [5,91]. SORT1 was proposed as the gene driving this association due to its role in enhancing LDL-C catabolism as reported in murine knockdown studies [92]. However, due to the inherent infrequency of rare variants, they are unlikely to exist in LD with common SNPs (tested in GWAS) in surrounding genes that may otherwise contribute to CAD/MI manifestation. Moreover, rs599839 was found to be localized to an expression quantitative trait locus (eQTL) (i.e. a region influencing gene expression levels) which was determined to regulate expression levels for all three genes (CELSR2, PSRC1, SORT1) in hepatocytes [93]. An additional SNP localized to the 3' UTR of CELSR2 (rs7528419) was independently associated with increased Lipoprotein-associated phospholipase A2 activity, which results in production of pro-inflammatory factors (mainly lysophospholipids) within the arterial wall ($P = 1.3x10^{-17}$; beta = 0.035 for effect allele) [94]. This SNP is also independently associated with prevalent CAD, which may be mediated by its aforementioned pro-inflammatory effects ($P=1.97x10^{-23}$) [94]. Given that this SNP is found within the 3' UTR of CELSR2, it may result in perturbed translational efficiency or mRNA stability which exhibit analogous functional effects to rare disease-causing variants within the coding sequence. A SNP found within transcript sequence of a given gene may also be less likely to functionally effect neighboring genes, namely PSRC1

and SORT1. Taken together, there is reason to confirm the role for the burden of rare disease-causing variants in CELSR2 for predisposing risk to MI.

5.3.4.2 APOA5

APOA5 acts as a co-activator to the LPL enzyme, which is responsible for hydrolyzing triglycerides localized within chylomicrons and VLDL [2,32]. The nominal association between burden of rare, disease-causing variants in APOA5 and MI in our work represents a replication of results from an exome sequencing study published by Do *et al.* 2016 [2], who also used the EOMI cohort samples (among others) in their analysis to identify association between rare variant burden and early MI. The association signal in our work was driven by 7 T1 nonsynonymous SNVs predicted to be damaging using the M-CAP *in silico* pathogenicity tool across 12 heterozygous carriers and a single nonsense variant across 1 heterozygous carrier (Table 5.2). A similar magnitude of effect for the association of APOA5 with CAD/MI was determined between our work and results from the Do *et al.* study (OR 2.72 and 2.2, respectively). The rs964184 SNP within the ZNF259-APOA5-APOA1 locus has consistently demonstrated genome-wide significant association with CAD/MI in GWAS meta-analyses [5]. However, the results in the aforementioned study in addition to our work identifies APOA5 as the risk gene driving association. The association of APOA5 with CAD/MI has also been shown to be mediated by elevated plasma triglycerides [2,95], consistent with its molecular function.

The demarcation of two known CAD/MI genes establishes a proof of concept for our case-only study design. Nevertheless, due to lack of individual-level data for controls,

it is essential to manually examine the biological functional of novel genes exhibiting significant or suggestive association in order to reconcile phenotypic correlates to the burden of rare variation.

Table 5.1: Association of CELSR2 and APOA5 with early MI in the EOMI European population.

Gene	EOMI European allele count	ExAC NFE CMAF	OR (95% CI)	P-value
CELSR2 Ψ	78	0.022945	3.08 (2.54-4.13)	$1.1x10^{-17}$
APOA5 †	13	0.00431674	2.72 (1.45-4.84)	0.001

Ψ Association signal observed using only T5 alleles; † Association signal observed using T1 alleles

5.3.5 Additive analysis identifies novel, biologically relevant genes in EOMI Europeans and Africans using the ExAC NFE and AFR populations as the control dataset

Several genes with biological significance to CAD/MI were identified in our case-only approach at either exome-wide or nominal significance. In EOMI Europeans, we identified nominal associations with endothelin-converting enzyme 2 (ECE2) (P = 0.002) and matrix metalloproteinase 9 (MMP9) (P = 0.0004) (Table 5.2). For EOMI Africans, we identified an exome-wide significant association with intestinal alkaline phosphatase (ALPI) ($P = 2.2x10^{-5}$) and a nominal association with hydroxy-delta-5-steroid dehydrogenase, 3 Beta- Steroid Delta-Isomerase 7 (HSD3B7) (P = 0.0001) (Table 5.2).

5.3.5.1 ECE2

ECE2 encodes for endothelin converting enzyme 2 which catalyzes the proteolytic conversion of big endothelin-1 to endothelin-1 within endothelial cells [96]. While the role of endothelin-1 in CAD/MI has been well documented, the mechanism by which big endothelin-1 could predispose individuals to disease remains elusive. However, recent epidemiological analyses have demonstrated positive correlation of increased levels of plasma big endothelin-1 with future CVD events (including MI and revascularization procedures) and lower event-free survival (P=0.016) [97,98], thus establishing it as a prospective prognostic marker. Furthermore, elevated big endothelin-1 has been shown to be associated with increased coronary calcification [99], which could represent a putative mechanism by which CAD/MI risk is conferred. The association signal for ECE2 was found to be driven by 18 T1 nonsynonymous SNVs across 26 heterozygous carriers (Table 5.2). All variants were predicted to be damaging by M-CAP.

5.3.5.2 MMP9

Matrix metalloproteinases (MMPs) are endopeptidases whose members are mostly responsible for the catalytic cleavage of proteins comprising the ECM - which are necessary for the formation of the fibrous cap overlaying atherosclerotic plaques [100]. MMPs are highly expressed in a wide-array of vascular cell types (i.e. endothelial cells and vascular smooth muscle cells) and have shown to be vital mediators of CAD/MI onset in both epidemiological and functional studies [100–102]. Over 20 subtypes of MMPs have been characterized with each having preferential affinity for different ECM proteins [100] which

has led investigators to suspect that this enzymatic family contains members that could either confer risk or protection to CAD/MI. MMP9 has consistently been shown to confer protection against plaque rupture by promoting vascular smooth muscle cell migration [101], which leads to reinforcement of plaque stability through increased ECM deposition. Moreover, murine models of MMP9 knockout ($MMP9^{-/-}$) have shown that MMP9 inhibits platelet aggregation by catalyzing the cleavage of fibrin within the ECM [103]. This is supported through clinical observations of increased MMP9 levels in individuals suffering an ACS (i.e. NSTEMI, STEMI), but not in individuals with stable CAD [104]. This is primarily due to a compensatory mechanism in which MMP9 is upregulated to heal ruptured plaques by promoting smooth muscle cell migration [101]. Taken together, these results support a protective role for MMP9 in plaque rupture and thrombus generation. Indeed, it is of high interest that we have discovered enrichment of rare disease-causing variants within MMP9 in an early MI cohort, as opposed to a case population that does not necessarily have to be afflicted with an ACS for study inclusion. Overall, we identify 6 T1 nonsynonymous SNVs in MMP9 across 19 heterozygous carriers (Table 5.2). All variants were predicted to be damaging by M-CAP.

5.3.5.3 ALPI

The ALPI gene is most highly expressed by duodenal enterocytes and encodes intestinal alkaline phosphatase, an enzyme localized to the intestinal brush-border membrane responsible for the detoxification (via dephosphorylation) of local lipopolysaccharides (LPS) [105]. LPS are able to induce a pro-inflammatory environment that

sufficiently enhances gut permeability and consequently enhances intestinal lipid absorption [106]. Therefore, ALPI represents a robust negative regulator of intestinal fat absorption. In fact, *in vivo* murine models have shown that ALPI$^{-/-}$ mice present with significantly elevated plasma triglycerides after exposure to high-fat diets [106]. We identified a total of 5 T5 nonsynonymous SNVs predicted to be damaging by M-CAP and a single nonsense variant across 9 heterozygous carriers (Table 5.2).

5.3.5.4 HSD3B7

The HSD3B7 gene demonstrates high expression in hepatocytes and encodes for 3 beta-hydroxysteroid dehydrogenase type 7, which participates in the enzymatic cascade responsible for converting cholesterol to bile acids (cholic acid and chenodeoxycholic acid) [107]. Specifically, HSD3B7 catalyzes a reduction reaction that converts 7α-hydroxycholesterol to 7α-hydroxy-4-cholesen-3-one, which is the second enzymatic process in the cholesterol to bile acid conversion [107]. Individuals homozygous for HSD3B7 variants typically develop congenital bile acid synthesis defect type I, which is characterized by impaired intestinal fat absorption [107]. Heterozygous carriers of HSD3B7 have also demonstrated markedly reduced enzymatic function and increased hepatic LDL-C [107], however, it remains to be determined as to whether this is can elevate plasma LDL-C to levels sufficient for atherosclerotic plaque development. A total of 5 T1 nonsynonymous SNVs predicted to be damaging by MCAP contributed to the association signal for HSD3B7 and were identified across 6 heterozygous carriers (Table 5.2).

Table 5.2: Association of biologically relevant genes with early MI in EOMI European or African populations.

Gene	EOMI European/African allele count	ExAC NFE/AFR CMAF	OR (95% CI) ǂ	P-value	EOMI effect population
ECE2 Ψ	26	0.012	1.90 (1.28-2.95)	0.002	European
MMP9 Ψ	19	0.0070	2.44 (1.48-3.95)	0.0004	European
ALPI †	9	0.0051	6.10 (2.64-12.7)	$2.2x10^{-5}$	African
HSD3B7 Ψ	6	0.0027	7.74 (2.54-18.7)	0.0001	African

Ψ Association signal determined with T5 alleles; † Association signal determined with T1 alleles

5.3.6 Additive and recessive analyses identifies novel, biologically relevant genes in DECODE cohort using a weighted estimate of all ExAC populations as the control dataset

We have identified 5 novel genes (3 additive and 2 recessive) enriched for rare disease-causing variants within DECODE. Under the additive model, we show nominal associations with Carcinoembryonic antigen-related cell adhesion molecule 1 (CEACAM1) (P=0.0002), Myotubularin Related Protein 9 (MTMR9) (P = 0.007), and DExH-Box Helicase 34 (DHX34) (P = 0.007). Under the recessive model, we identify an exome-wide significant association with butyrophilin like 3 (BTNL3) (P = 0.0002) and a nominal association with interleukin 7 receptor (IL7R) (P = 0.001).

5.3.6.1 CEACAM1

The CEACAM1 gene demonstrates high expression in both vascular endothelial cells and hepatocytes. CEACAM1 encodes for a transmembrane adhesion molecule which

has been shown to regulate endothelial cell permeability and mediate numerous intracellular signaling cascades [108]. *In vivo* murine knockout of CEACAM1 (CEACAM1$^{-/-}$) exhibit a variety vascular aberrations [109] that produce appropriate conditions for the development of CAD. Specifically, CEACAM1$^{-/-}$ mice exhibit 2-fold greater protein expression of VCAM-1 [109], which participates in leukocyte adhesion to the endothelial monolayer, leading to the establishment of an inflammatory foci. This observation provides support for a potential role of CEACAM1 in negatively regulating VCAM-1 expression. Furthermore, CEACAM1$^{-/-}$ mice show vast reductions (60%) in endothelial nitric oxide content [109], which was shown to compromise vessel relaxation and result in hypertension. Lastly, endothelial cells of CEACAM1 $^{-/-}$ mice are characterized by aberrant intercellular junctions due to decreased expression of vascular endothelial cadherin (VE-cadherin), resulting in increased vessel permeability [108,109].

In other work, CEACAM1 $^{-/-}$ mice developed spontaneous atherosclerosis and were characterized by significantly elevated levels of total cholesterol, LDL-C and triglycerides [110] in addition to all the vascular abnormalities discussed above. This evidence is corroborated in our work as the 2 carriers of CEACAM1 variants demonstrated elevated levels of LDL-C and/or triglycerides beyond the 90th percentiles (Table 5.3). The association signal for CEACAM1 was driven by a single frameshift indel in two heterozygous carriers (Table 5.4). Neither individual was an FH or FCH mutation carrier.

Further functional evidence is necessary to confirm a causal role of this variant is CAD development and progression. However, given that the variant in question is

heterozygous and disruptive, it is highly likely that the native CEACAM1 function as at least 50% compromised.

Table 5.3: LDL-C and triglyceride levels for CEACAM1 rare variant carriers.

DECODE ID	LDL-C (mmol/L)	Triglycerides (mmol/L)
0008	3.2	3.0
44	6.2	3.0

5.3.6.2 MTMR9

The MTMR9 gene encodes Myotubularin-related protein 9 which demonstrates high expression in peripheral blood mononuclear cells. The function of MTMR9 has not been well documented, but SNPs within this gene have previously demonstrated nominal associations with systolic blood pressure in GWAS [111]. More recently, SNPs within MTMR9 were found to significantly associate with hypertension and impaired fasting glucose in a Japanese population [112]. Several GWAS meta-analyses have also consistently demonstrated strong association between loci regulating vessel tone and CAD, such as SH2B3 ($P=1.0x10^{-9}$) and NOS3 ($P=1.7x10^{-7}$) [5]. Given the substantial risk conferred by hypertension on CAD onset, we sought to identify whether carriers of rare disease-causing variants in MTMR9 presented with elevated systolic blood pressure as compared to non-carriers. After adjustment for age, sex, BMI, and anti-hypertensive medication (adding 10 mmHg to systolic blood pressure), we indeed confirmed a significant association ($P = 9.6x10^{-4}$, beta = 27.6 mmHg) and will look to replicate this finding in additional cohorts. Overall, a total of 3 T1 nonsynonymous SNVs were identified across 5 heterozygous carriers (Table 5.4).

5.3.6.3 DHX34

DHX34 is a member of the DExH/D box gene family and exhibits high expression within endothelial cells. Members of the DExH/D box gene family play essential roles in regulating mRNA expression by participating in mRNA synthesis, ribosomal biogenesis and nonsense-mediated decay (NMD) [113]. Of these processes, DHX34 has shown to be necessary for the activation of NMD in human cells [114]. NMD is a surveillance mechanism that functions to identify and degrade mRNA harbouring premature stop codons, thereby preventing the translation of truncated proteins [114,115]. Variants perturbing the native function of DHX34 may result in disruption of numerous cellular processes which could confer risk for complex disease. We sought to investigate endothelial-specific effects of rare DHX34 variants via reprogramming peripheral blood leukocytes into iPSCs and subsequently differentiating these cells to the endothelial lineage using glycogen synthase kinsase-3 inhibitors and vascular endothelial growth factor.

Endothelial cells derived from DHX34 variant carriers have demonstrate perturbed tube formation compared to healthy iPSC-derived endothelial cells and human umbilical vein endothelial cells (HUVECs). This observation is suggestive of differential angiogenic capacity between DHX34-carrier and healthy endothelial cells which likely results in decreased ability to form compensatory vascular networks to perfuse ischemic tissue. Secondly, DHX34-carrier endothelial cells demonstrate increased expression (compared to baseline established by healthy iPSC-derived endothelial cells and HUVECs) of I-CAM1 on their cell surface during inactive states. I-CAM1 is an adhesive proteins responsible for mediating the process of leukocyte extravazation (see section 1.8.2) into the *tunica intima*

and establishing the foci for plaque formation. Therefore, expression of I-CAM1 in an inactive state can render endothelial cells primed for an inflammatory processes. Lastly, we observe that DHX34-carrier endothelial cells are characterized by an aberrant "jagged" morphology, which impedes their ability to form adhesive monolayers necessary to regulate cell-to-cell permeability. We observe a total of 3 T1 nonsynonymous SNVs predicted to be damaging by M-CAP across 3 heterozygous carriers (Table 5.4).

5.3.6.4 BTNL3

BTNL3 is a member of the butyrophilin superfamily and demonstrates high expression within the antigen-presenting cells (APCs) of the innate immune system [116]. BTNL3 functions as a potent negative regulator of T-cell activation as T-cells have been robustly shown to play a pivotal role in atherosclerotic plaque instability [40,116]. Specifically, *in vitro* models of cultured murine $CD4^+$ T-cells demonstrate significant inhibition in proliferative capacity when co-cultured with both APCs that overexpress BTNL3 and stimulatory ligands [116]. Interestingly, addition of a BTNL3 monoclonal antibody was able to rescue normal T-cell activation as demonstrated by a return to baseline proliferation rate in the presence of stimulatory ligands [116]. We identified a single homozygous carrier of a T5 nonsynonymous SNV in BTNL3 predicted to be deleterious according to SIFT and damaging according to Polpyhen2-HDIV/HVAR (Table 5.4).

5.3.6.5 IL7R

IL7R is largely expressed in both B and T-lymphocytes and is responsible for their maturation and proliferation as IL7R knockout ($IL7R^{-/-}$) mice show a 10-fold reduction in

precursor B-cells [117]. However, it has also been shown that IL7R is expressed in human microvascular endothelial cells (HMEC) where it shown to positively contribute to endothelial cell proliferation in a dose-dependent manner after the administration of exogenous Interleukin 7 (IL-7) [118]. We identified a single homozygous carrier of a T5 nonsynonymous SNV in IL7R with a CADD score >20 (Table 5.4).

Table 5.4: Association of biologically relevant genes with early CAD in DECODE.

Gene	DECODE allele or carrier count	ExAC wCMAF or wMACF	OR (95% CI)	P-value	Association model
CEACAM1 †	2	0.0002	95.4 (10.2-417)	0.0002	Additive
MTMR9 †	5	0.0070	4.15 (1.40-10.7)	0.007	Additive
DHX34 †	3	0.0051	7.75 (1.61-24.2)	0.007	Additive
BTNL3 Ψ	1	$4.91x10^{-6}$	-	0.0002	Recessive
IL7R Ψ	1	$2.4x10^{-5}$	-	0.001	Recessive

Ψ Association signal determined with T5 alleles; † Association signal determined with T1 alleles; wMACF represents the DECODE-weighted frequency of the CMACF for a given gene (see section equation 5.2 and section 5.2.10)

5.3.7 SKAT analysis identifies CLEC4D as novel CAD gene in DECODE cohort using CVD-free ORIGIN samples as the control dataset

Unlike burden test, variance-component tests such as SKAT are able to regress the CAD phenotype on each rare variant within a gene in addition any specified covariates [23]. As such, individual variants are not assumed to act in the same direction or confer the same magnitude of effect. Therefore, we used a less stringent pathogenicity filter (T1 alleles that are disruptive or nonsynonymous SNV predicted to be deleterious/damaging by SIFT or

Polphen2-HDIV/HVAR) to conduct SKAT, which has substantially greater power (compared to burden tests) when neutral or protective variants are expected to be present.

After adjusting for age, sex, and BMI, we discovered an exome-wide significant association with C-Type Lectin Domain Family 4 Member D (CLEC4D) ($P = 3.1x10^{-6}$), which is highly expressed within cells of the innate immune system, including macrophages, monocytes and neutrophils [119]. *In vivo* analysis of murine knockout models (i.e. $CLEC4D^{-/-}$) was found to induce pulmonary inflammation mediated by excessive neutrophil recruitment [119]. While the role for CLEC4D has not yet been demonstrated in CAD, the molecular phenotype exhibited in knockout studies is indeed consistent with what would be expected to confer CAD risk. Therefore, this gene represents a candidate for functional follow-up in the macrophage arm of this study.

5.4 Conclusion

In this chapter, we have leveraged the size of the ExAC database to identify numerous genes that associate with CAD at exome-wide or nominal significance using the in aggregate burden test for rare variants. Associations identified in previously known CAD genes (i.e. CELSR2 and APOA5) using the EOMI case cohort represent a proof-of-concept of our case-only study design. Moreover, the delineation of novel, biologically relevant genes in EOMI provides grounds to perform replication analyses in larger case cohorts in order to robustly establish their putative roles in conferring risk for CAD. We do not report any exome-wide significant findings in the DECODE cohort under an additive model of inheritance, but have identified several genes demonstrating both nominal significance and

biological plausibility in the context of CAD. One such example is DHX34, for which we have established a robust vascular phenotype that is consistent with what is observed during CAD progression. We have indeed identified 2 genes: BTNL3 and IL7R that reach exome-wide and nominal significance, respectively, under the recessive model of inheritance. Since associations for both genes are driven by variants with homozygous or trans compound heterozygous genotypes, they functionally represent human gene knockouts. Consequently, the demarcation of biologically relevant genes in recessive-based analyses should be followed-up for clinical validation, regardless of the statistical significance level. Lastly, we used ~400 internal CVD-free control samples to conduct a SKAT analysis which detected an exome-wide significant association with CLEC4D.

We expect that the continual growth of the DECODE study will facilitate the discovery of additional genes associated with CAD and provide further supporting evidence genes currently exhibiting nominal significance.

CHAPTER 6 – 'N of 1' Benchmarking for the Calibration of Individual Sequences to Big Data: A Novel Methodology to Facilitate Construction of Rare Variant Gene Risk Scores

6.1 Introduction

The recent development of large sequencing datasets such as ExAC and gnomAD have substantially improved the ability to discern rare disease causing variants of high effect from benign polymorphisms in cohorts with small sample sizes. Recent work published by Wilfert *et al.* 2016 [120] has also demonstrated utility of these 'Big Data' sets in facilitating the identification of genome-wide significant variants in single samples (i.e. the 'N of 1' problem) by leveraging null distributions generated from gene-based pathogenicity scores. However, this approach demonstrates best results in cases of Mendelian disease and may not necessarily be well suited for stratifying individuals according to risk for complex diseases such as CAD, which can be driven by the additive effect of multiple rare risk alleles across several genes. Moreover, these approaches cannot necessarily account for biases inherent within population sub-structure and sequencing artefacts that may be confounding 'N of 1' association signals.

In this chapter we establish rationale that population sub-structure and sequencing artifacts can readily impede the ability to detect 'N of 1' associations. By using consensus reference sequences that are devoid of artefactual variants, we demonstrate that individual ethnic background significantly compromises single case sequences from achieve calibration with Big Data sets. Additionally, we use high coverage sequences with variants filtered at multiple stringencies to demonstrate the impact of sequencing artifacts on single case calibration and the corresponding gains/losses achieved in sensitivity and specificity.

We propose the development of single case correction factors (CF), which are coefficients weighted based on the magnitude of discrepancy in total mutation load

between local sequences and Big Data. We demonstrate practical applications for CF in the calibration of single case exomes to the ExAC dataset (i.e. 'N of 1' benchmarking) for the construction of rare variant gene scores (RVGS), which will allow for un-confounded assessment of individual-level associations with CAD based on the count of variant alleles (meeting specified pathogenicity criteria) in well-established CAD risk genes. RVGS may be incorporated into risk stratification models and be used as a factor to inform appropriate clinical decisions. Furthermore, single sample RVGS can be leveraged to compute cohort-wide associations, which is particularly attractive for populations of small sample size that maybe underpowered detect rare variant association signals using standard approaches.

6.2 Methods

6.2.1 DECODE study population

The DECODE population is described in detail in section 2.3.1.

6.2.2 Leuven study population

A total of 77 CAD-free individuals of Belgian ancestry were recruited as part of a severe white-matter disease study by the University of Leuven underwent whole-exome sequencing in our Laboratory using methods described in sections 2.4.3-5. This cohort was used as an internal reference population for downstream analyses to ascertain potential biases inherent to our sequencing and bioinformatics pipelines.

6.2.3 Variant calling and annotation

Variant calling and annotation was performed as described in sections 2.4.7-8.

6.2.4 Variant pathogenicity filtering

Variant pathogenicity filtering was conducted identically as described for association analysis (section 5.2.7). Briefly, T5 alleles that were either disruptive or nonsynonymous SNVs predicted deleterious/damaging by SIFT or Polyphen2-HDIV/HVAR were used to generate gene-based cumulative sum plots for GIAB reference samples. T5 alleles that were either disruptive or nonsynonymous SNVs predicted to be damaging by M-CAP were used for the rest of the analyses.

6.2.5 Benchmarking correction factors for local comparison sequences

Correction factors for each sample were calculated based on the ratio of mutation loads between a comparison sample and the ExAC dataset. Specifically, variant counts across ~18500 autosomal protein-coding genes were aggregated to generate **observed** variant counts in a comparison sample, and then compared to the aggregate **expected** aggregate variant count obtained from ExAC frequencies across the same set of genes (equation 6.1)

$$CF_i = \frac{\sum_{j=1}^{M} V_{ij}}{\sum_{j=1}^{M} V_{ExACj}} \qquad \text{Equation 6.1}$$

where CF represents the correction factor for individual i which is equal to the ratio between the aggregate variant count V for gene j to M. Expected ExAC variant counts were generated using ethnic-specific frequencies dictated by the ethnicity of the

comparison sample (e.g. samples of European ancestry would generate aggregate variant counts based on the Non-Finnish European subgroup in ExAC) in order to prevent population stratification bias.

Variant counts were determined only across exons that did not demonstrate differential coverage between the comparison sample and ExAC in order to ensure that there was equal sensitivity to detect variants. The difference in % 20X coverage metric (defined in section 2.4.12) was calculated for each exon (Ei in equation 6.2) in ~18500 protein-coding genes. Exons exhibiting an absolute difference in % 20X coverage of > 10% were excluded and thus did not contribute to mutation load for the comparison sample or ExAC (conditional 1).

$$\left.\begin{array}{l} if \mid \Delta\ \%\ 20X\ coverage \mid_{Ei} > 10 \\ \quad then\ exclude; \\ \quad otherwise\ keep \end{array}\right\} \text{ Conditional 1}$$

Lastly, all variant counts used to generate the correction factor were standardized to the minor allele according to conditional 2. Briefly, frequencies corresponding to the alternate allele (i.e. alternate allele frequency (AAF)) that were greater than 0.5 were flipped to the minor allele. This procedure serves to circumvent the exclusion of variants based on arbitrary assignment of reference and alternate alleles.

$$\left.\begin{array}{l} if\ AAF > 0.5; then \\ \quad then\ MAF = 1 - AAF; \\ \quad otherwise\ MAF = AAF \end{array}\right\} \text{ Conditional 2}$$

6.2.6 Benchmarking correction factors for GIAB consensus sequences

In order to evaluate the accuracy of exome-wide mutation load in ExAC, we determined correction factors using high-confidence sequence variants characterized by the GIAB consortium (using methodology developed by Zook *et al.* 2014 [121]) for 2 reference samples of differing ethnicities: NA12878 (European) and NA24631 (East Asian). Variants called in these reference sets were harmonized across 5 sequencing technologies, 7 read mappers and 3 variant callers to generate a "consensus" variant callset that were used as benchmarks against the ExAC dataset. Defining consensus variant calls also allows for the identification of "difficult-to-sequence" regions across the genome largely due to the presence of segmental duplications, short tandem repeats and structural variations. Characterization of these regions for different reference samples can be used to generate "high-confidence" genomic intervals that are free of difficult to sequence regions. High confidence region files and variant call sets (in variant call format) for version 3.2.2 of NA12878 and version 3.3 of NA24631 were obtained from the GIAB ftp repository (ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/NA12878_HG001/NISTv3.3.2/) and (ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/ChineseTrio/HG005_NA24631_son/NISTv3.3/) for use in our benchmarking analysis.

Correction factors were generated as in Equation 6.1. Variant counts were determined only across exons that were both present within NIST v.3.2.2 high confidence regions and exhibiting high coverage in ExAC (% 20X coverage > 90) (conditional 3). All variants contributing to the correction factor were also standardized to the minor allele as stated in conditional 2.

$$\left.\begin{array}{l} if\ (\%\ 20X\ coverage < 90\ in\ ExAC)_{Ei}\ or\ (outsite\ high\text{-}confidence\ region)_{Ei} \\ then\ exclude; \\ otherwise\ keep \end{array}\right\}$$ Conditional 3

6.2.7 NA12878 & NA24631 Ampliseq

Genomic DNA for the NA12878 and NA24631 reference samples was obtained from Coriell Cell Repositories [122] and underwent Ion Ampliseq™ library and template preparation as described in section 2.4.3. Exome libraries underwent template preparation on 2 IonChef™ instruments and were subsequently exome sequenced on 2 Ion S5XL™ sequencers (4 total runs), respectively as described in sections 2.4.4-5. This combination approach was selected in order to minimize technical biases that may have been exclusive to a particular instrument. Binary alignment (BAM) files generated from all 4 sequencing runs were merged using the samtools *merge* function to generate a combined BAM file with a mean coverage depth of 700X for NA12878 and 660X for NA24631.

6.2.8 Evaluating true positives, false positives and false negatives in the NA12878-Ampliseq sequence at different variant filtering stringencies

Variant calls for the high-depth NA12878-Ampliseq sequence were filtered according to default TVC 5.2 variant filtering settings in addition to the low, medium, and high Damiati variant filtering stringencies (discussed in section 2.4.7). Variants generated across all filtering stringencies were evaluated against the NA12878 GIAB truth set in order to ascertain the proportion of calls that were true positives, false positives and false negatives. The proportion of variant calls mutually present in both NA12878 Ampliseq variants and

the truth set were evaluated as true positives. Conversely, variant calls present within NA12878 Ampliseq, but absent in the gold-standard set were set as false positives. Lastly, calls present in the gold-standard set, but absent in NA12878-Ampliseq were evaluated as false negatives. Identical variants with different genotypes (e.g. heterozygous in NA12878-Ampliseq but homozygous in the gold-standard) were considered discordant and would be evaluated as either false positives or false negatives. All metrics were computed using RTG Tools' *vcfeval* tool [123].

6.2.9 Mutation load visualization

Plots representing the cumulative sum of variant counts meeting specified frequency and pathogenicity criteria are shown to demonstrate the contribution of each gene to the overall mutation load. Gene indices on the plots (x-axis) are ordered by gene-based mutation rate (i.e. gene-based variant count adjusted for total gene length). In order to facilitate interpretation of these plots, an annotated example is provided below. The plateau of each cumulative sum distribution represents the mutation load.

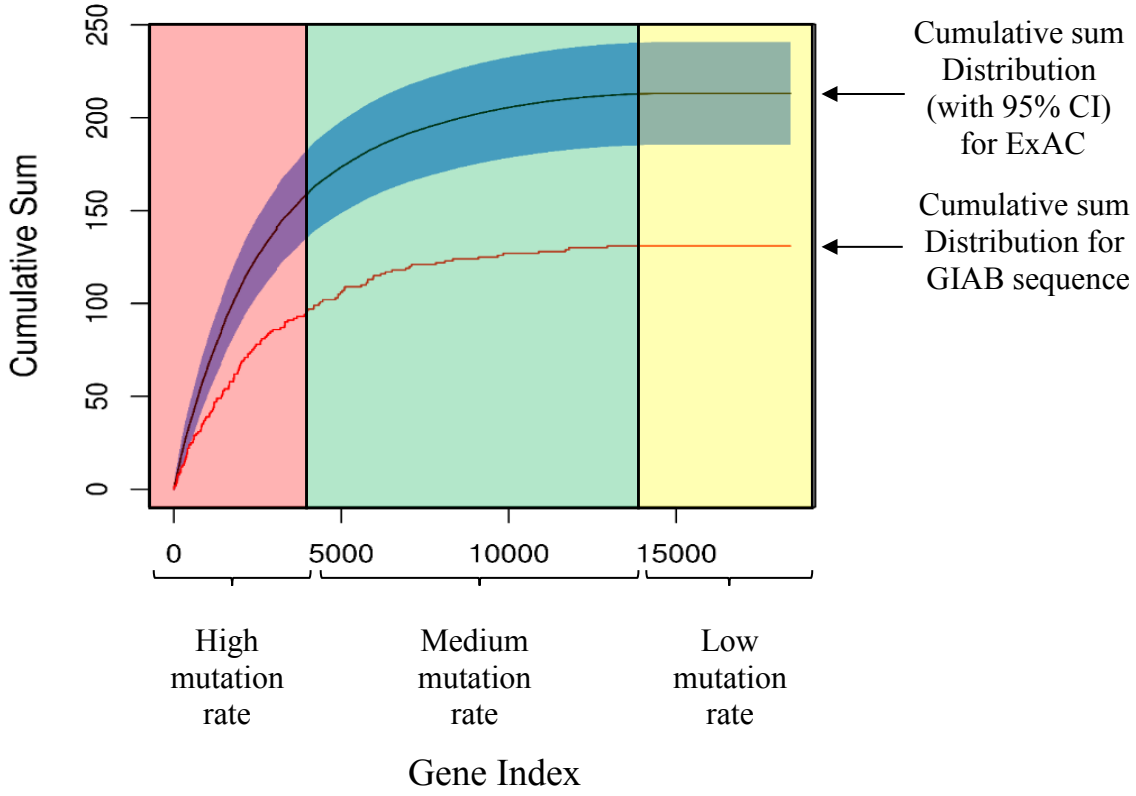


Figure 6.1: Annotation figure for cumulative sum distribution plots. Red, green, and yellow shaded areas represent genes with high, medium, and low mutation rate in ExAC, respectively. Discrepancy in slope between ExAC (solid black line) and GIAB sequence (solid red line) indicate the difference in contribution of each gene to the overall

6.2.10 Statistical analysis for cumulative sum distributions and mutation loads

The 95% confidence intervals for ExAC cumulative sum distributions was determined by computing the cumulative Bernoulli variance on ExAC gene-based CMAF (equation 6.2).

$$95\%\ CI = 2\,p_j \pm 1.96\sqrt{\sum_{j=1}^{M} 2\,p_j(1-p_j)} \qquad \text{Equation 6.2}$$

where p represents the ExAC CMAF from genes j to M and 2 represents a coefficient to convert the CMAF to a cumulative minor allele count (CMAC).

P-values for differences in correction factors between or within ethnic groups was determined using a student's t-test. All computations for this section and remaining methods sections were conducted in R version 3.2.2 unless otherwise stated.

6.2.11 Rare variant gene scores

Gene-based odds ratios generated from rare variant association analysis between the EOMI cohort and ExAC were used to compute beta coefficients according to equation 6.3.

$$\beta_j = \ln(OR)_j \qquad \text{Equation 6.3}$$

Each gene-based beta value was subsequently used to weight the aggregate rare variant burden score (i.e. total number of rare alleles meeting a pre-specified pathogenicity criteria) observed for the corresponding gene within each individual from the DECODE (cases) and Leuven (control) cohorts. Weighted burden scores were summed across all genes to generate a rare variant gene score per individual according to equation 6.4

$$RVGS_i = \sum_{j=1}^{M} \beta_j RVB_{ij} \qquad \text{Equation 6.4}$$

where $RVGS$ represents the rare variant gene score for individual i across genes j to M and where RVB represents the aggregate rare variant burden score for individual i in gene j.

6.2.12 Statistical analysis for RVGS

Per-sample rare variant gene scores were adjusted for correction factors and converted to standardized z-scores using population means and standard deviation for RVGSs generated from ExAC. Standard deviation of ExAC RVGS was computed through a beta-coefficient weighted cumulative Bernoulli variance based on the ExAC CMAF across a pre-specified set of genes (equation 6.5)

$$\sigma_{ExAC} = \sqrt{\sum_{j=1}^{M} 2\,\beta_j^2 p_j(1-p_j)} \qquad \text{Equation 6.5}$$

where p represents the ExAC CMAF from genes j to M and 2 represents a coefficient to convert the CMAF to a CMAC. 95% CI for proportions of individuals with significant RVGS was calculated using the modified Wald method.

6.2.13 Endothelial secretome

Due to the recent demarcation of numerous loci involved in vessel wall biology being associated with CAD, we sought to determine whether proteins encompassing the endothelial secretome demonstrated an increased mutational load in early CAD patients relative to CAD-free controls using our 'N of 1' benchmarking method. We used PUBMED literature searches to identify recent publications that may have curated a list of secreted proteins that characterize endothelial cells. A study published by Tunica *et al.* 2009 mapped a total of 71 proteins to human umbilical vein endothelial cells which were used in our analysis (Supplementary Table 8). Gene identifiers for all 71 proteins were obtained from the HUGO Gene Nomenclature Committee (HGNC) database.

6.3 Results

6.3.1 Evaluating the effect of population structure on mutation load using consensus (GIAB) sequences

After intersecting exons through "high-confidence" genomic regions and with sufficiently high coverage in ExAC (i.e. % 20X coverage > 90), an overall mutation load of 215 and 211 was determined for ExAC's Non-Finnish European (NFE) and East Asian (EAS) populations, respectively (Figure 6.2 and Table 6.1). Corresponding mutation loads for the GIAB European (NA12878-GIAB) and East Asian (NA24631-GIAB) reference samples were 131 and 174, resulting in correction factors (CF) of 0.61 and 0.83, respectively (Figure 6.2 and Table 6.1). The cumulative sum distributions for ExAC NFE and NA12878-GIAB are significantly discrepant ($P<0.05$) across genes that are highly mutable (as indicated by the difference between the slopes) (Annotated Figure and Figure 6.2A). In contrast, ExAC EAS frequencies adhere closely to observed variant counts in the NA24631-GIAB reference samples across all genes (Figure 6.2B).

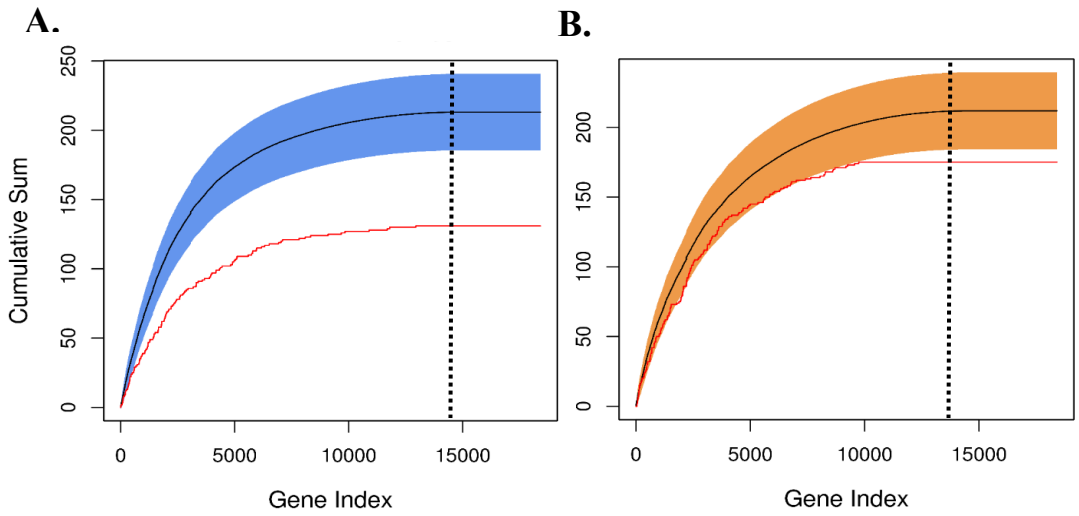


Figure 6.2: Gene-based cumulative sum distributions for ExAC and GIAB reference samples. (A) Black and red solid lines represent the ExAC NFE population and the NA12878-GIAB reference sample, respectively. **(B)** Black and red solid lines represent the ExAC EAS population and the NA24631-GIAB reference sample, respectively. Shaded areas correspond to 95% confidence intervals on ExAC variant counts. Gene indices (x-axis) are ordered from highest to lowest ExAC mutation rate. Genes beyond the black dotted line are mutually absent of variants in ExAC and reference samples.

Table 6.1: Total mutation loads and corresponding correction factors for GIAB sequences and ExAC for NFE and EAS ancestries.

	ExAC NFE	NA12878-GIAB	ExAC EAS	NA24631-GIAB
Mutation load	215	131	211	175
CF	0.61		0.83	

In order to identify whether the patterns observed for T5 alleles in the NFE and EAS populations was consistent across the entire allele frequency distribution, all variants meeting a pre-specified pathogenicity criteria were spit across 6 allele frequency bins (Table 6.2). Cumulative sum distributions were also plotted on an exon-by-exon basis in order to reduce stochasticity observed in the gene-based plots by increasing the total

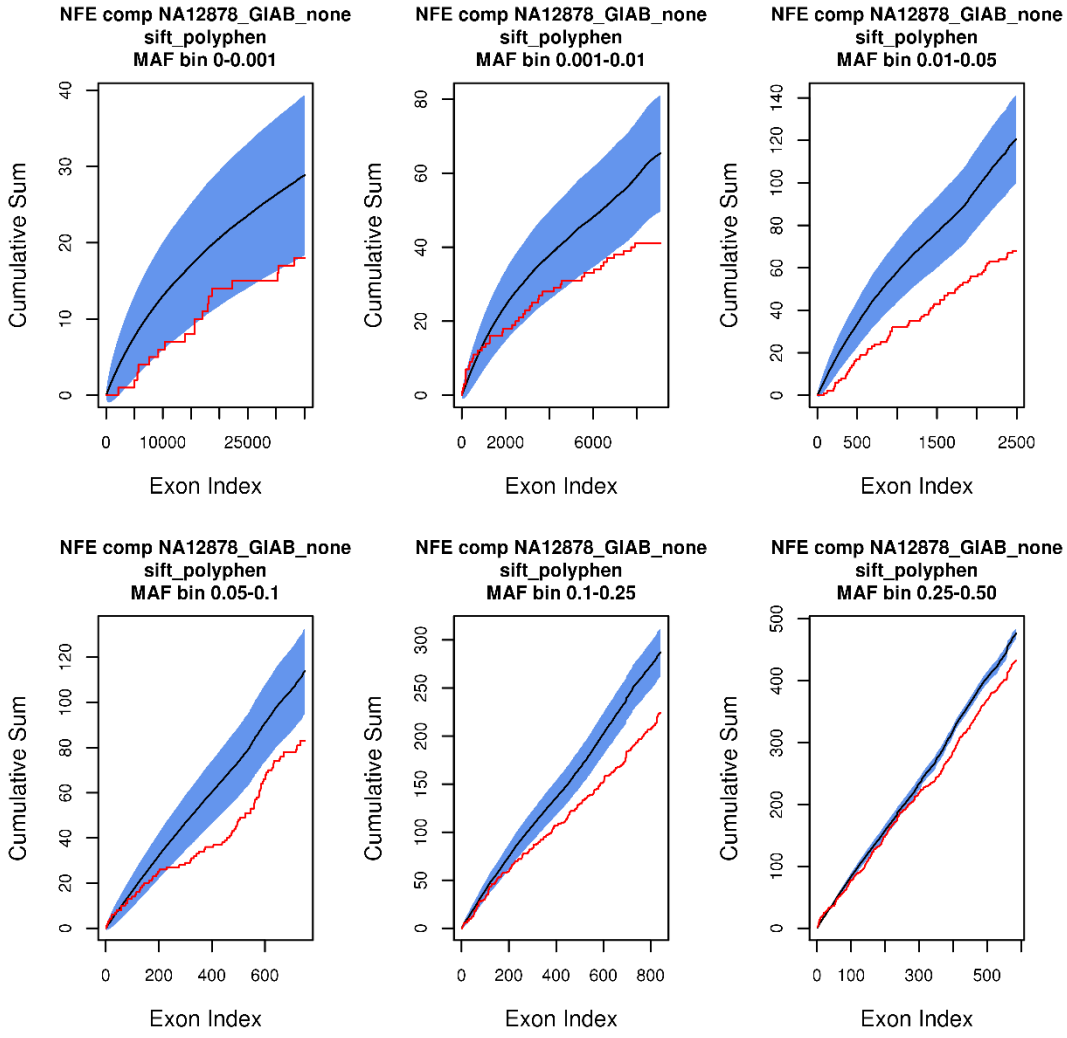
number of observations and refine interpretation of population-specific effects (Figure 6.2A and 6.2B).

The mean CFs determined for the "rare" allele frequency bins (i.e. bins 1-3) were 0.81 +/- 0.12 and 0.61 +/- 0.03 for ExAC EAS-NA24631 GIAB and ExAC NFE-NA12878 GIAB, respectively (Table 6.3). Mean CFs for "common" allele frequency bins (i.e. bins 4-6) were 1.00 +/- 0.05 and 0.80 +/- 0.09 (Table 6.3). The CFs for rare allele frequency bins were significantly more impactful (i.e. further from CF of 1) than those for common bins in the ExAC NFE-NA12878 GIAB comparison ($P<0.05$). No significant difference between rare and common CFs in the ExAC EAS-NA24631 GIAB comparison was observed. However, CFs were significantly higher and better calibrated (i.e. CF closer to 1) across all allele frequency bins in ExAC EAS-NA24631 GIAB compared to ExAC NFE-NA12878 GIAB ($P<0.05$).

Table 6.2: Allele frequency bins used to generate mutation loads

Bin number	Lower bound frequency	Upper bound frequency
1	0	0.001
2	0.001	0.01
3	0.01	0.05
4	0.05	0.1
5	0.1	0.25
6	0.25	0.50

A.



B.

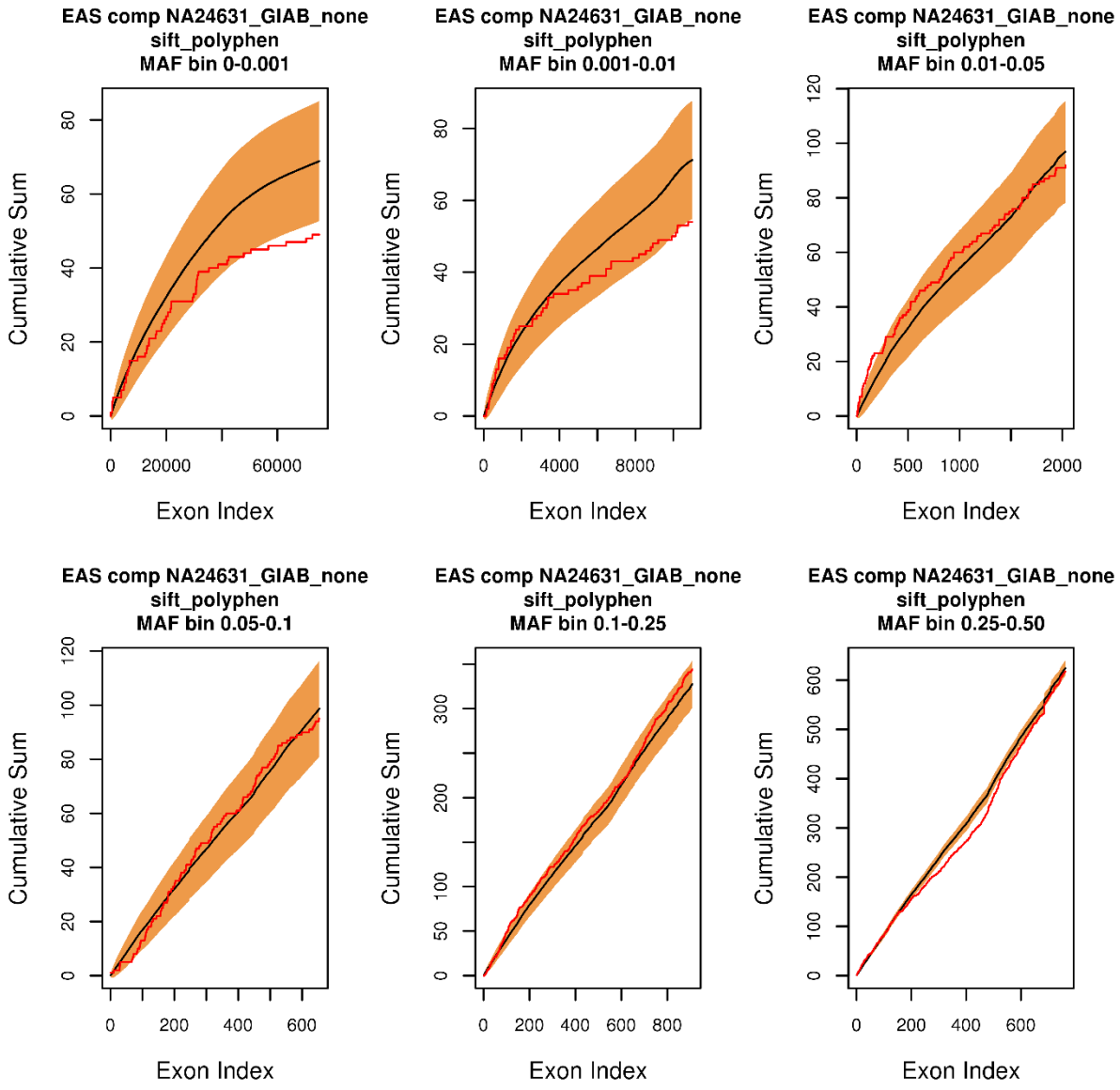


Figure 6.3: Exon-based cumulative sum distributions for ExAC and GIAB reference samples across 6 allele frequency bins. (A) Black and red solid lines represent the ExAC NFE population and the NA12878-GIAB reference sample, respectively. **(B)** Black and red solid lines represent the ExAC EAS population and the NA24631-GIAB reference sample, respectively. Shaded areas correspond to 95% confidence intervals on ExAC variant counts. Exon indices (x-axis) are ordered from highest to lowest ExAC mutation rate.

Table 6.3: Correction factors for ExAC EAS-NA24631 GIAB and ExAC NFE-NA12878 GIAB. Blue and green sections represent rare and common allele frequency bins, respectively. Means +/- SD CFs are provided for rare and common bins in each comparison.

Bin number	Correction factor ExAC EAS and NA24631 GIAB	Mean +/- SD	Correction factor ExAC NFE and NA12878 GIAB	Mean +/- SD
1	0.71	0.81 +/- 0.12	0.62	0.61 +/- 0.03
2	0.76		0.63	
3	0.95		0.56	
4	0.96	1.00 +/- 0.05	0.73	0.80 +/ 0.09
5	1.05		0.78	
6	1.00		0.91	

6.3.2 Evaluating the effect of sequencing artefacts on mutation load using the NA12878-Ampliseq sequence

Variant calls for the NA12878-Ampliseq sequence were filtered using default TVC 5.2 parameters in addition to all 3 Damiati filters (see section 6.2.7 in Methods and section 2.4.7) in order to evaluate the effects of filtering stringency on CFs. Default stringency resulted a mutation load closest to what is expected in ExAC (173/215; CF = 0.8) (Table 6.4), but retains the most false positive variants (Table 6.5). The mutation loads steadily deviate away from what is expected in ExAC as variant filtering stringency increases, resulting in more extreme CFs (Table 6.4).

Table 6.4: Mutation loads and CFs for NA12878-Ampliseq sequence using Damiati variant filtering criteria

	Mutation load	**CF**
ExAC NFE	215	NA
NA12878-Ampliseq Default	173	0.80
NA12878-Ampliseq Low	159	0.74
NA12878-Ampliseq Med	156	0.72
NA12878-Ampliseq High	143	0.66

Table 6.5: Proportion of variants retained (relative to default settings on TVC 5.2) for each Damiati stringency using T5 alleles that are either disruptive or nonsynonymous SNVs predicted to be deleterious/damaging according to SIFT or Polyphen2-HDIV/HVAR. True positives, false positives, and false negatives were determined using the NA12878-GIAB as the truth set

Variant filtering stringency	**True Positive retained (%)**	**False positives lost (%)**	**False negatives gains (%)**
Low	99	35	4
Medium	97	40	23
High	87	54	44

6.3.3 Evaluating heterogeneity in mutation loads of local European sequences

We generate cumulative sum plots, mutation loads, and CFs for 39 DECODE Europeans and 77 Leuven sequences using T5 alleles that were either disruptive or nonsynonymous SNVs evaluated as "damaging" by M-CAP. Using this pathogenicity criteria, the mean mutation loads for DECODE and Leuven were 77.4 +/- 12.8 and 72.5 +/- 12.7, respectively. The corresponding CFs were both determined to be 1.03 +/- 0.15. No significant differences were detected for either mutation load or CF between DECODE

and Leuven cohorts (P>0.05). Overall, 15% (95% CI 6.9-30.1%) and 45% (95% CI 34.8-56.5%) of CFs were < 1 for DECODE and Leuven, respectively (Figure 6.3).

Cumulative sum distribution plots for exemplar local samples showcasing variability in mutation load are provided in Figure 6.4. The three samples depicted illustrate instances where mutation load is inflated (Figure 6.5A), deflated (Figure 6.5B), and well-calibrated (Figure 6.5C) as compared to what is expected in ExAC. The mutation loads and corresponding correction factors for each interrogated DECODE and Leuven sample is provided in Supplementary Table 9.

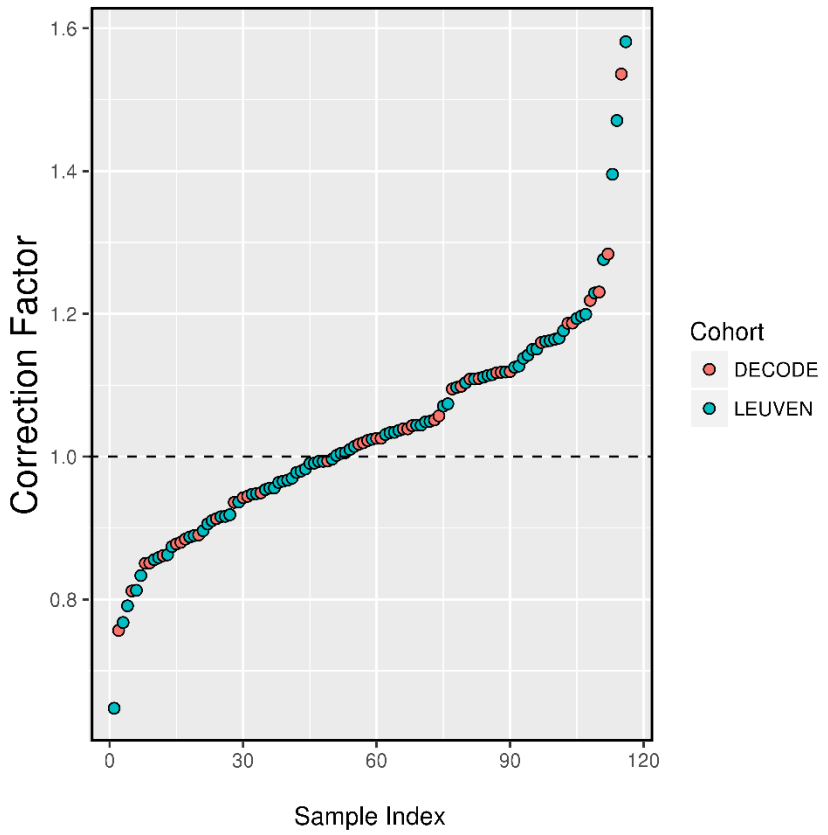


Figure 6.4: Scatter plot of correction factors for each sample of DECODE (orange) and Leuven (blue) cohorts. Correction factors are ordered from smallest **to** largest.

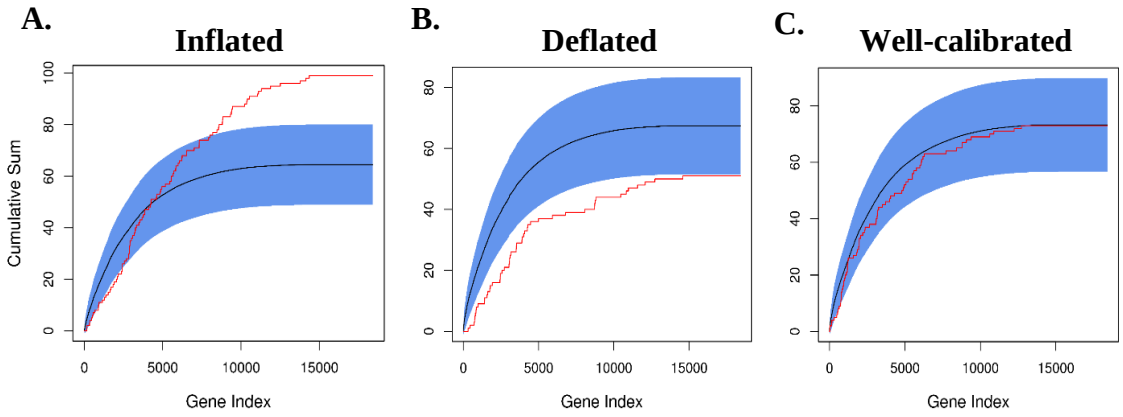


Figure 6.5: Examples cumulative sum distributions generated from local sequences showcasing different single-sample calibration with ExAC. Plots depict single sample sequences that that are inflated **(A)**, deflated **(B)**, and well-calibrated **(C)**

6.3.4 Using CFs to calibrate construction of RVGS in CAD and CAD-free cohorts

The gene-based beta values obtained from association analysis conducted using EOMI European cases and ExAC NFE controls (see section 5.3.4) were used to construct sample-wise RVGS in DECODE Europeans and Leuven (CAD-free) participants according to equations 5.2 and 5.3. Several p-value thresholds were used to arbitrate the selection of which genes from the above association analysis contributed to the RVGS calculation (Table 6.5). Using a stringent p-value threshold of 0.001, we observed a significantly higher RVGS in DECODE European participants compared to expected mean in ExAC NFE (0.240 vs 0.130; z-score = 1.75; p-value = 0.04) (Table 6.6). In contrast, no significant deviation in mean RVGS was determined for CAD-free controls as compared to expectation in ExAC NFE (0.09 vs 0.130; z-score = -0.721; p-value = 0.231) (Table 6.6). In increasing the association gene-based p-value threshold by 10-fold (i.e. $p < 0.01$), we again observed significant elevation in mean RVGS among DECODE European

participants (0.83 vs. 0.60; z-score = 1.84; p-value = 0.03) and no significant deviation in mean RVGS among CAD-free controls (0.46 vs. 0.60; z-score = -1.61; p-value = 0.053) compared to expectation in ExAC NFE (Table 6.6). ORs for CAD were also calculated based on RVGS in DECODE (OR 1.12 for $p<0.001$ and OR 1.27 for $p<0.01$) (Table 6.6). Among all p-value thresholds that were interrogated, CAD-free controls were determined to have mean z-scores < 0 (i.e. consistent depletion of CAD-risk alleles) while DECODE Europeans had z-scores > 0 (i.e. consistent enrichment of CAD-risk alleles).

Before incorporation of correction factors and in using $p < 0.01$, 13% (95% CI 5-27%) (5/39) DECODE Europeans and 5% (95% CI 2-13%) (4/77) CAD-free control participants demonstrated significant enrichment of CAD-risk alleles (i.e. z-score > 1.64). Following incorporation of correction factors, 15% (95% CI 7-30%) (6/39) DECODE and 3% (95% CI 0.2-10%) (2/77) CAD-free participants were enriched for CAD-risk alleles. Summary of z-scores and CAD ORs for $p < 0.01$ criteria are provided in Supplementary Table 10 for DECODE and CAD-free participants. The RVGS for a single DECODE participant (bolded sample in Supplementary Table 10) was substantially increased after incorporation of the CF, resulting in gain of an 'N of 1' association signal (i.e. z-score > 1.64; p-value < 0.05).

Table 6.6: Association of CF-adjusted RVGS in DECODE Europeans and CAD-free controls

EOMI European vs. ExAC NFE p-value threshold	N risk Genes used Ψ	ExAC mean RVGS	ExAC RVGS SD	DECODE z-score of mean RVGS	CAD-free z-score of mean RVGS	P-value DECODE	P-value CAD-free	CAD OR in DECODE
0.001	13	0.13	0.28	1.75	-0.72	0.04	0.231	1.12
0.01	85	0.60	0.54	1.84	-1.61	0.03	0.053	1.27

Ψ Refers to number of genes from EOMI-ExAC association analysis used to build RVGS

6.4 Discussion

Benchmarking of human sequences to Big Data represents a novel approach to accurately and confidently identify enrichment of risk alleles that may confer susceptibility to complex diseases such as CAD. This is especially attractive for the conduct of 'N of 1' association analyses where the validity of results is largely dependent on the degree of calibration between a local sequence and large sequencing databases (i.e. ExAC). We demonstrate the rationale for the development of sample-specific correction factors in order to calibrate the mutation load of a single sample to what is expected in large sequencing databases. Specifically, in leveraging highly accurate sequencing genotypes from reference samples of differing ethnicity, we show the impact of population structure in producing deviations from expected mutation loads.

The ExAC NFE population demonstrated significantly elevated mutation load as compared to the NA12878 GIAB European sample across all genes (regardless of mutation rate) whereas the ExAC EAS population was well-calibrated with the NA24631 GIAB East Asian reference sample across all genes. It is unlikely that the difference in mutation rate

between NA12878 GIAB and ExAC NFE is due to technical artefacts in the ExAC database because if this were the case, the East Asian NA24631 sample should also be similarly affected by artefacts and exhibit an elevated mutation load.

Instead, the variation in degree of calibration appears to be population-specific and likely attributable to differences in population sub-structure within Europe and East Asia. Since Europeans have demonstrated a high degree of continental migration [124] and are characterized by greater geographical heterogeneity, it is likely that variation in population sub-structure is accounting for the difference in mutation loads between ExAC NFE and NA12878 GIAB. In contrast to Europeans, East Asian populations have exhibited significantly less continental migration [125]and are localized to homogenized geographical regions. This is reflected by similar mutation loads observed between ExAC EAS and NA24631 GIAB.

In an effort to understand whether the magnitude of difference or similarity in mutation load was being driven by rare or common variants, we stratified the analysis according to 6 MAF bins. It was observed that mean correction factors calculated based on the mutation loads for ExAC NFE-NA12878 GIAB comparison were significantly more impactful (i.e. higher-weighted) for rare frequency bins compared to common. This observation is as expected since it has been well established that rare variants exhibit a higher degree of geographical specificity compared to common variants [14,126]. No significant difference was observed between rare and common frequency bins in the ExAC EAS-NA24631 GIAB comparison, which further reflects the degree of population homogeneity in East Asian populations.

We also show the impact of sequencing artefacts on mutation load calibration. In using the NA12878-Ampliseq high coverage sample, there was increased deviation from expected mutation load in ExAC NFE as variant filtering stringency increased. However, an increased stringency substantially dampened false positive rate while also increasing the frequency false-negative variants.

Due to the degree of genetic variability influenced by differences in population structure, sequencing chemistries, variant calling algorithms, and variant filtering criteria, we postulate that the incorporation of a correction factor can significantly democratize the conduct of 'N of 1' association analyses and provide robust results that are not otherwise confounded.

In the last portion of this chapter, we establish practical applications of per-sample correction factors for the construction of RVGS, which have the ability to detect enrichment of disease-causing alleles of high effect on the individual (i.e. 'N of 1') and cohort level (mean 'N of 1'). Since the calculation of RVGS is based on summary statistics (i.e. odds ratios) obtained from large association analyses, they can be leveraged by smaller cohorts to detect meaningful enrichment of risk alleles from multiple genes which **1)** cannot be detected in standard association analysis and **2)** can inform clinical decisions for individual samples exhibiting significant elevation in their RVGS compared to expectation.

We performed an arbitrary selection of gene-based p-value thresholds obtained the EOMI European and ExAC NFE association to select genes used to construct RVGS for DECODE and internal CAD-free controls using ExAC to generate the null distribution. Using a p-value threshold of 0.001, we detected significantly elevated mean RVGS in

DECODE compared to ExAC expectation. Mean RVGS in CAD-free controls did not deviate significantly from ExAC expectation, which demonstrates that elevation in DECODE was not due to technical artifacts inherent in our sequencing or bioinformatics pipeline. When the p-value threshold increased to 0.01, we again observed significant elevation of mean RVGS in DECODE while mean RVGS in CAD-free controls remained insignificant. We tested several intermediary p-value thresholds which all trended in enrichment (positive z-scores) and depletion (negative z-scores) of CAD risk alleles in DECODE and CAD-free controls, respectively. Not all thresholds resulted in significant findings, which corroborates the necessity to simulate various scenarios of variant pathogenicity, frequency, and p-value thresholds in order to arrive at a consensus for optimal thresholds that are able to effectively discriminate between significant and null findings.

Lastly, we demonstrate the "gain of significance" for a single DECODE sample after incorporation of correction factors to the RVGS. While mean RVGS remained significant for DECODE Europeans on the cohort level, 1 individual was found to have a deflated mutation load compared to ExAC NFE, which resulted in a higher-weighted correction factor (bolded sample in Supplementary Table 10). Although this is a small proportion, these findings are proof-of-concept for the necessity of calibrating individual samples to large sequencing databases in order to facilitate correct interpretation of 'N of 1' association analyses which will have implications for potential clinical management in terms of risk stratification.

CHAPTER 7 – Conclusion and Future Directions

7.1 Conclusion

In this work, we sought to validate the hypothesis that rare, protein-altering variants confer risk for the development of EOCAD. To evaluate this, we formulated three primary objectives: **1)** to identify and biologically characterize rare, protein-altering genetic mutations responsible for very early CAD using burden and variance component testing under case-only and case-control study designs, **2)** to develop per-sample correction factors that can be used to calibrate 'N of 1' benchmarking analyses in practical applications such as rare variant association testing and calculation of rare variant gene scores, and **3)** to determine the prevalence of Mendelian dyslipidemias (especially familial hypercholesterolemia) in young, angiographically-proven CAD patients by evaluating rare, protein-altering variants in known genes.

The first objective involved conducting RVAS in the form of burden and variance component testing in order to detect genes that were enriched for putative disease-causing variants beyond what is expected due to chance. Rare variant burden testing was conducted with two case populations (the EOMI cohort and DECODE) in a case-only study design using ethnically-matched populations in the ExAC dataset as controls. We replicated associations in two known CAD genes: CELSR2 and APOA5 at exome-wide and nominal significance, respectively. Emergence of these genes as top associations in addition to several genes with strong biological rationale for CAD/MI (ECE2, MMP9, ALPI, and HSD3B7) demonstrates the efficacy of our case-only study design and bioinformatics pipeline. In DECODE (using both burden and variance component testing), we identified exome-wide and nominal associations of several genes involved in endothelial and immune

cell function, suggesting prominent role of these cell types in the development of very early CAD.

To complete our second objective, we selected several genes from the EOMI vs. ExAC analysis that exhibited nominal statistical significance to construct RVGS that could be leveraged to conduct single sample (i.e. 'N of 1') association analyses in DECODE participants by using derived descriptive statistics from public sequencing consortia (i.e. ExAC) as the reference population. However, we established that both population structure and frequency of sequencing false positives/negatives rigorously impacts cumulative mutation burden and, in turn, the degree of calibration between single samples and ExAC. Consequently, we developed single sample CFs to calibrate mutation loads in order to ameliorate spurious 'N of 1' associations driven by differences in population structure and sequencing artefacts as opposed to the additive effect of rare variants. This process, termed 'N of 1' benchmarking, identified 6 DECODE participants with significant enrichment of CAD/MI risk alleles. Additionally, we observed significant enrichment of CAD/MI risk alleles across all DECODE participants (cohort-level), but not in a CAD-free population, which was processed with identical sequencing and bioinformatics pipelines. We conclude that CF-adjusted RVGS can facilitate detection of individual and cohort-level enrichment of risk alleles beyond what is detectable by standard RVAS methods. Moreover, CF-adjusted RVGS can be used to inform disease prognosis and risk stratification, which can be especially useful for young individuals with advanced history of CAD/MI.

To satisfy our third objective, we generated a semi-automated pipeline to delineate variants conferring risk for Mendelian dyslipidemias (especially FH). Briefly, our pipeline

involved the selection of rare protein-altering variants across 24 genes in the WDLV that have been previously annotated as "pathogenic" or "likely pathogenic" in the ClinVar database. The lipid panel and family pedigree of individuals harbouring such variants were subsequently evaluated to determine overall penetrance and degree of familial co-segregation. Variants with no ClinVar annotation or established as VUS were manually assessed in InterVar by adjusting ACMG/AMP criteria based on patient phenotype, family history/co-segregation, mode of inheritance, and evidence of established functional effects in *in vitro/in vivo* models. Variants that generated a "pathogenic", "likely pathogenic", or "VUS" annotation in InterVar were established as causal, assuming phenotypic penetrance/family history was present. In using this pipeline, we identified 3 cases of FH and 2 cases of FCH, resulting in an overall diagnostic yield of 10%.

In order to assess whether very early CAD populations harbour an increased frequency of FH-causing mutations, we conducted an association analysis using count variables from CAD-free and unselected patients populations as control data, which was mined from the MIGen and CHARGE consortia. We observed a significant enrichment of FH-causing mutations in DECODE both before and after applying an LDL-C cutoff for variant inclusion. Moreover, we observed an enrichment of FH-causing mutations even compared to other premature CAD populations with more liberal age-cutoffs than those applied in DECODE. Our findings provide support for the application of systematic genetic screening for individuals with very early CAD in order to evaluate the presence of putative disease-causing mutations within FH genes. This is especially important considering

individuals harbouring these mutations may not be receiving appropriate clinical management or may circumvent clinical FH diagnosis due being on statin regimens.

7.2 Future directions

The large majority of our future work will be building upon objective 2 in this work. Although we successfully demonstrated application of the CF-adjusted RVGS on the individual and cohort-level, we must conduct extensive simulation on random variables in order to assess impact of incorporation of the CF on statistical metrics including estimated OR and statistical power. Specifically, scenarios inclusive of deviations in population structure and frequency of sequencing artefacts between single exomes and ExAC will be incorporated into simulation models in order to comprehensively assess their individual on the above metrics. We will also be looking to generate more robust CAD/MI RVGS with the use of additional case samples from both the HeartGO cohorts of the NHLBI GO ESP6500 and the MIGen Exome Sequencing Consortium. Implementation of additional case samples will provide increased statistical power to effectively detect gene-based association signals driven by rare variants. Lastly, we hope to use our methodology to construct RVGS for other complex diseases (e.g. type II diabetes mellitus, stroke) and for gene-set panels (e.g. the endothelial secretome). Such applications will provide the opportunity to demonstrate the utility of our RVGS across a variety of complex diseased states in addition to demarcating specific gene-sets involved in complex disease development. Moreover, it will provide means to help stratify individual prognostic risk for a variety of conditions in order to reduce disease incidence.

References

1. Mozaffarian D, Benjamin EJ, Go AS, et al. Executive Summary: Heart Disease and Stroke Statistics-2016 Update: A Report From the American Heart Association. *Circulation*. 2016;133(4):447-454. doi:10.1161/CIR.0000000000000366.

2. Do R, Stitziel NO, Won H-H, et al. Exome sequencing identifies rare LDLR and APOA5 alleles conferring risk for myocardial infarction. *Nature*. 2015;518(7537):102-106. doi:10.1038/nature13917.

3. Vranckx P, Leebeek FWG, Tijssen JGP, et al. Peri-procedural use of rivaroxaban in elective percutaneous coronary intervention to treat stable coronary artery disease. The X-PLORER trial. *Thromb Haemost*. 2015;114(2):258-267. doi:10.1160/TH15-01-0061.

4. Sayols-Baixeras S, Lluís-Ganella C, Lucas G, Elosua R. Pathogenesis of coronary artery disease: focus on genetic risk factors and identification of genetic variants. *Appl Clin Genet*. 2014;7:15-32. doi:10.2147/TACG.S35301.

5. the CARDIoGRAMplusC4D Consortium. A comprehensive 1000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat Genet*. 2015;47(10):1121-1130. doi:10.1038/ng.3396.

6. Mangino M, Spector T. Understanding coronary artery disease using twin studies. *Heart*. 2013;99(6):373-375. doi:10.1136/heartjnl-2012-303001.

7. Yusuf S, Hawken S, Ounpuu S, et al. Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case control study. *Lancet*. 2004;364(9438):937-952. doi:10.1016/S0140-6736(04)17018-9.

8. Lloyd-Jones DM, Nam B-H, D'Agostino RB, et al. Parental cardiovascular disease as a risk factor for cardiovascular disease in middle-aged adults: a prospective study of parents and offspring. *JAMA*. 2004;291:2204-2211. doi:10.1001/jama.291.18.2204.

9. Marenberg ME, Risch N, Berkman LF, Floderus B, de Faire U. Genetic susceptibility to death from coronary heart disease in a study of twins. *N Engl J Med*. 1994;330(15):1041-1046. doi:10.1056/NEJM199404143301503.

10. Stitziel NO, MacRae CA. A clinical approach to inherited premature coronary artery disease. *Circ Cardiovasc Genet*. 2014;7(4):558-564. doi:10.1161/CIRCGENETICS.113.000152.

11. Naoumova RP, Tosi I, Patel D, et al. Severe hypercholesterolemia in four British families with the D374Y mutation in the PCSK9 gene: Long-term follow-up and treatment response. *Arterioscler Thromb Vasc Biol*. 2005;25(12):2654-2660. doi:10.1161/01.ATV.0000190668.94752.ab.

12. Deloukas P, Kanoni S, Willenborg C, et al. Large-scale association analysis identifies new risk loci for coronary artery disease. *Nat Genet*. 2013;45(1):25-33. doi:10.1038/ng.2480.

13. Gorlov IP, Gorlova OY, Sunyaev SR, Spitz MR, Amos CI. Shifting Paradigm of Association Studies: Value of Rare Single-Nucleotide Polymorphisms. *Am J Hum Genet*. 2008;82(1):100-112. doi:10.1016/j.ajhg.2007.09.006.

14. Tennessen J a, Bigham AW, O'Connor TD, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*. 2012;337(6090):64-69. doi:10.1126/science.1219240.

15. Nelson MR, Wegmann D, Ehm MG, et al. An Abundance of Rare Functional Variants in 202 Drug Target Genes Sequenced in 14,002 People. *Science (80-).* 2012;337(6090):100-104. doi:10.1126/science.1217876.

16. Jeff JM, Peloso GM, Do R. What can we learn about lipoprotein metabolism and coronary heart disease from studying rare variants ? 2016:1-6. doi:10.1097/MOL.0000000000000277.

17. Bamshad MJ, Ng SB, Bigham AW, et al. Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet*. 2011;12(11):745-755. doi:10.1038/nrg3031.

18. Do R, Kathiresan S, Abecasis GR. Exome sequencing and complex disease: Practical aspects of rare variant association studies. *Hum Mol Genet*. 2012;21(R1). doi:10.1093/hmg/dds387.

19. gnomAD browser | genome Aggregation Database. http://gnomad.broadinstitute.org/.

20. Exome Aggregation Consortium. http://exac.broadinstitute.org.

21. Lek M, Karczewski KJ, Minikel E V., et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. 2016;536(7616):285-291. doi:10.1038/nature19057.

22. Li B, Leal S. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet*. 2008;83:311-321. doi:10.1016/j.ajhg.2008.06.024.

23. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet*. 2011;89(1):82-93. doi:10.1016/j.ajhg.2011.05.029.

24. Khera A V., Won H-H, Peloso GM, et al. Association of Rare and Common Variation in the Lipoprotein Lipase Gene With Coronary Artery Disease. *JAMA*. 2017;317(9):937. doi:10.1001/jama.2017.0972.

25. Lange LA, Hu Y, Zhang H, et al. Whole-exome sequencing identifies rare and low-frequency coding variants associated with LDL cholesterol. *Am J Hum Genet*. 2014;94(2):233-245. doi:10.1016/j.ajhg.2014.01.010.

26. Auer PL, Nalls M, Meschia JF, et al. Rare and Coding Region Genetic Variants Associated With Risk of Ischemic Stroke: The NHLBI Exome Sequence Project. *JAMA Neurol*. 2015;72(7):781-788. doi:10.1001/jamaneurol.2015.0582.

27. Lohmueller KE, Sparsø T, Li Q, et al. Whole-exome sequencing of 2,000 Danish individuals and the role of rare coding variants in type 2 diabetes. *Am J Hum Genet*. 2013;93(6):1072-1086. doi:10.1016/j.ajhg.2013.11.005.

28. Kathiresan S. Developing medicines that mimic the natural successes of the human genome: Lessons from NPC1L1, HMGCR, PCSK9, APOC3, and CETP. *J Am Coll Cardiol*. 2015;65(15):1562-1566. doi:10.1016/j.jacc.2015.02.049.

29. Investigators TMIGC. Inactivating Mutations in NPC1L1 and Protection from Coronary Heart Disease. *N Engl J Med*. 2014;371(22):2072-2082. doi:10.1056/NEJMoa1405386.

30. Heart N. Loss-of-Function Mutations in APOC3, Triglycerides, and Coronary Disease. *N Engl J Med*. 2014:1-10. doi:10.1056/NEJMoa1307095.

31. Dewey FE, Gusarova V, O'Dushlaine C, et al. Inactivating Variants in ANGPTL4 and Risk of Coronary Artery Disease. *N Engl J Med*. 2016;374(12):1123-1133. doi:10.1056/NEJMoa1510926.

32. Naukkarinen J, Ehnholm C, Peltonen L. Genetics of familial combined hyperlipidemia. *Curr Opin Lipidol*. 2006;17(3):285-290. doi:10.1097/01.mol.0000226121.27931.3f.

33. Libby P. Inflammation in atherosclerosis. *Arterioscler Thromb Vasc Biol*. 2012;32(9):2045-2051. doi:10.1161/ATVBAHA.108.179705.

34. Libby P, Theroux P. Pathophysiology of coronary artery disease. *Circulation*. 2005;111(25):3481-3488. doi:10.1161/CIRCULATIONAHA.105.537878.

35. Ambrose JA, Singh M. Pathophysiology of coronary artery disease leading to acute coronary syndromes. *F1000Prime Rep*. 2015;7:08. doi:10.12703/P7-08.

36. Liao JK. Linking endothelial dysfunction with endothelial cell activation. *J Clin Invest*. 2013;123(2):540-541. doi:10.1172/JCI66843.

37. Heitzer T, Schlinzig T, Krohn K, Meinertz T, Munzel T. Endothelial Dysfunction, Oxidative Stress, and Risk of Cardiovascular Events in Patients With Coronary Artery Disease. *Circulation*. 2001;104(22):2673-2678. doi:10.1161/hc4601.099485.

38. Crea F, Liuzzo G. Pathogenesis of acute coronary syndromes. *J Am Coll Cardiol*. 2013;61(1):1-11. doi:10.1016/j.jacc.2012.07.064.

39. Li ZY, Howarth SPS, Tang T, Gillard JH. How critical is fibrous cap thickness to carotid plaque stability? A flow-plaque interaction model. *Stroke*. 2006;37(5):1195-1199. doi:10.1161/01.STR.0000217331.61083.3b.

40. Libby P. Mechanisms of Acute Coronary Syndromes and Their Implications for Therapy. *N Engl J Med*. 2013;368(21):2004-2013. doi:10.1056/NEJMra1216063.

41. Barnett IJ, Lee S, Lin X. Detecting Rare Variant Effects Using Extreme Phenotype Sampling in Sequencing Association Studies. *Genet Epidemiol*. 2013;37(2):142-151. doi:10.1002/gepi.21699.

42. Peloso GM, Rader DJ, Gabriel S, Kathiresan S, Daly MJ, Neale BM. Phenotypic extremes in rare variant study designs. *Eur J Hum Genet*. 2016;24(6):924-930. doi:10.1038/ejhg.2015.197.

43. Thormaehlen AS, Schuberth C, Won HH, et al. Systematic Cell-Based Phenotyping of Missense Alleles Empowers Rare Variant Association Studies: A Case for LDLR and Myocardial Infarction. *PLoS Genet*. 2015;11(2):1-23. doi:10.1371/journal.pgen.1004855.

44. McKenna A, Hanna M, Banks E, et al. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20(9):1297-1303. doi:10.1101/gr.107524.110.

45. Damiati E, Borsani G, Giacopuzzi E. Amplicon-based semiconductor sequencing of human exomes: performance evaluation and optimization strategies. *Hum Genet*. 2016;135(5):499-511. doi:10.1007/s00439-016-1656-8.

46. Genome in a Bottle Consortium (GIAB). https://www.nist.gov/programs-projects/genome-bottle.

47. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 2010;38(16):e164. doi:10.1093/nar/gkq603.

48. Liu X, Wu C, Li C, Boerwinkle E. dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. *Human Mutation*. 2016.

49. Jagadeesh KA, Wenger AM, Berger MJ, et al. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nat Genet*. 2016;48(12):1581-1586. doi:10.1038/ng.3703.

50. 1000 Genomes Project Consortium, Auton A, Brooks LD, et al. A global reference for human genetic variation. *Nature*. 2015;526(7571):68-74. doi:10.1038/nature15393.

51. EVS. Exome Variant Server. *NHLBI GO Exome Seq Proj*. 2014. http://evs.gs.washington.edu/EVS/.

52. PLINK 1.90 beta. https://www.cog-genomics.org/plink2.

53. Sherry ST. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*. 2001;29(1):308-311. doi:10.1093/nar/29.1.308.

54. Wang J, Raskin L, Samuels DC, Shyr Y, Guo Y. Genome measures used for quality control are dependent on gene function and ancestry. *Bioinformatics*. 2014;31(3):318-323. doi:10.1093/bioinformatics/btu668.

55. Cooper DN, Mort M, Stenson PD, Ball E V, Chuzhanova NA. Methylation-mediated deamination of 5-methylcytosine appears to give rise to mutations causing human inherited disease in CpNpG trinucleotides, as well as in CpG

dinucleotides. *Hum Genomics*. 2010;4(6):406-410. doi:10.1186/1479-7364-4-6-406.

56. Yang Y, Muzny DM, Reid JG, et al. Clinical Whole-Exome Sequencing for the Diagnosis of Mendelian Disorders. *N Engl J Med*. 2013;369(16):1502-1511. doi:10.1056/NEJMoa1306555.

57. Lee H, Deignan JL, Dorrani N, et al. Clinical Exome Sequencing for Genetic Identification of Rare Mendelian Disorders. *JAMA*. 2014;312(18):1880. doi:10.1001/jama.2014.14604.

58. Soutar AK, Naoumova RP. Mechanisms of disease: genetic causes of familial hypercholesterolemia. *Nat Clin Pract Cardiovasc Med*. 2007;4(4):214-225. doi:10.1038/ncpcardio0836.

59. Myocardial Infarction Genetics Consortium. http://www.kathiresanlab.org/collaborators/myocardial-infarction-genetics-exome-sequencing-consortium/.

60. CHARGE Consortium: Cohorts for Heart & Aging Research in Genomic Epidemiology. http://www.chargeconsortium.com/.

61. Khera A V., Won HH, Peloso GM, et al. Diagnostic Yield and Clinical Utility of Sequencing Familial Hypercholesterolemia Genes in Patients With Severe Hypercholesterolemia. *J Am Coll Cardiol*. 2016;67(22):2578-2589. doi:10.1016/j.jacc.2016.03.520.

62. Fu J, Kwok S, Sinai L, et al. Western database of lipid variants (WDLV): A catalogue of genetic variants in monogenic dyslipidemias. *Can J Cardiol*. 2013;29(8):934-939. doi:10.1016/j.cjca.2013.01.008.

63. Landrum MJ, Lee JM, Benson M, et al. ClinVar: Public archive of interpretations of clinically relevant variants. *Nucleic Acids Res*. 2016;44(D1):D862-D868. doi:10.1093/nar/gkv1222.

64. Li Q, Wang K. InterVar: Clinical Interpretation of Genetic Variants by the 2015 ACMG-AMP Guidelines. *Am J Hum Genet*. 2017. doi:10.1016/j.ajhg.2017.01.004.

65. Hegele RA, Ban MR, Hsueh N, et al. A polygenic basis for four classical Fredrickson hyperlipoproteinemia phenotypes that are characterized by hypertriglyceridemia. *Hum Mol Genet*. 2009;18(21):4189-4194. doi:10.1093/hmg/ddp361.

66. Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. *Nat Rev Genet*. 2006;7(2):85-97. doi:10.1038/nrg1767.

67. Schunkert H, König IR, Kathiresan S, et al. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nat Genet*. 2011;43(4):333-338. doi:10.1038/ng.784.

68. Coronary T, Disease A, Consortium G. A genome-wide association study in Europeans and South Asians identifies five new loci for coronary artery disease. *Nat Genet*. 2011;43(4):339-344. doi:10.1038/ng.782.

69. Brautbar A, Leary E, Rasmussen K, Wilson DP, Steiner RD, Virani S. Genetics of familial hypercholesterolemia. *Curr Atheroscler Rep*. 2015;17(4):491. doi:10.1007/s11883-015-0491-z.

70. Talmud PJ, Shah S, Whittall R, et al. Use of low-density lipoprotein cholesterol gene score to distinguish patients with polygenic and monogenic familial hypercholesterolaemia: A case-control study. *Lancet*. 2013;381(9874):1293-1301. doi:10.1016/S0140-6736(12)62127-8.

71. Wald DS, Bangash FA, Bestwick JP. Prevalence of DNA-confirmed familial hypercholesterolaemia in young patients with myocardial infarction. *Eur J Intern Med*. 2015;26(2):127-130. doi:10.1016/j.ejim.2015.01.014.

72. Al-Rasadi K, Al-Waili K, Al-Sabti HA, et al. Criteria for diagnosis of familial hypercholesterolemia: A comprehensive analysis of the different guidelines, appraising their suitability in the Omani Arab population. *Oman Med J*. 2014;29(2):85-91. doi:10.5001/omj.2014.22.

73. Goldstein JL, Brown MS. The LDL receptor. *Arterioscler Thromb Vasc Biol*. 2009;29(4):431-438. doi:10.1161/ATVBAHA.108.179564.

74. Fouchier SW, Dallinga-Thie GM, Meijers JCM, et al. Mutations in stap1 are associated with autosomal dominant hypercholesterolemia. *Circulation Research*. 2014.

75. Futema M, Shah S, Cooper JA, et al. Refinement of Variant Selection for the LDL Cholesterol Genetic Risk Score in the Diagnosis of the Polygenic Form of Clinical Familial Hypercholesterolemia and Replication in Samples from 6 Countries. *ClinChem*. 2014;(1530-8561 (Electronic)).

76. Wiesbauer F, Blessberger H, Azar D, et al. Familial-combined hyperlipidaemia in very young myocardial infarction survivors (≤40 years of age). *Eur Heart J*. 2009;30(9):1073-1079. doi:10.1093/eurheartj/ehp051.

77. Abul-Husn NS, Manickam K, Jones LK, et al. Genetic identification of familial hypercholesterolemia within a single U.S. health care system. *Science (80-)*. 2016;354(6319):aaf7000. doi:10.1126/science.aaf7000.

78. Robinson JG, Goldberg AC. Treatment of adults with familial hypercholesterolemia and evidence for treatment: Recommendations from the National Lipid Association Expert Panel on Familial Hypercholesterolemia. *J Clin Lipidol*. 2011;5(3 SUPPL.). doi:10.1016/j.jacl.2011.03.451.

79. Gerstein HC, Yusuf S, Riddle MC, Ryden L, Bosch J. Rationale, design, and baseline characteristics for a large international trial of cardiovascular disease prevention in people with dysglycemia: The ORIGIN Trial (Outcome Reduction with an Initial Glargine Intervention). *Am Heart J*. 2008;155(1). doi:10.1016/j.ahj.2007.09.009.

80. NHLBI GO-ESP: Early-Onset Myocaridal Infarction (Broad EOMI). https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000279.v2.p1.

81. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. *Bioinformatics*. 2010;26(22):2867-2873. doi:10.1093/bioinformatics/btq559.

82. Wigginton JE, Cutler DJ, Abecasis GR, Abecasis R, Abecasis GR. *A Note on Exact Tests of Hardy-Weinberg Equilibrium*. Vol 76. 2005:887-893. doi:10.1086/429864.

83. Danecek P, Auton A, Abecasis G, et al. The variant call format and VCFtools. *Bioinformatics*. 2011;27(15):2156-2158. doi:10.1093/bioinformatics/btr330.

84. O'Connell J, Gurdasani D, Delaneau O, et al. A General Approach for Haplotype Phasing across the Full Spectrum of Relatedness. *PLoS Genet*. 2014;10(4). doi:10.1371/journal.pgen.1004234.

85. Prof Jonathan Marchini Prof Goncalo Abecasis Prof Richard Durbin. Haplotype Reference Consortium. *http://www.haplotype-reference-consortium.org/*. 2014. http://www.haplotype-reference-consortium.org/.

86. R Developement Core Team. R: A Language and Environment for Statistical Computing. *R Found Stat Comput*. 2015;1:409. doi:10.1007/978-3-540-74686-7.

87. Westfall P, Wolfinger R. Multiple tests with discrete distributions. *Am Stat*. 1997;51(1):3-8. doi:10.2307/2684683.

88. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-variant association analysis: Study designs and statistical tests. *Am J Hum Genet*. 2014;95(1):5-23. doi:10.1016/j.ajhg.2014.06.009.

89. Samani NJ, Braund PS, Erdmann J, et al. The novel genetic variant predisposing to coronary artery disease in the region of the PSRC1 and CELSR2 genes on chromosome 1 associates with serum cholesterol. *J Mol Med*. 2008;86(11):1233-1241. doi:10.1007/s00109-008-0387-2.

90. Arvind P, Nair J, Jambunathan S, Kakkar V V., Shanker J. CELSR2-PSRC1-SORT1 gene expression and association with coronary artery disease and plasma lipid levels in an Asian Indian cohort. *J Cardiol*. 2014;64(5):339-346. doi:10.1016/j.jjcc.2014.02.012.

91. Willer CJ, Schmidt EM, Sengupta S, et al. Discovery and refinement of loci associated with lipid levels. *Nat Genet*. 2013;45(11):1274-1283. doi:10.1038/ng.2797.

92. Kjolby M, Andersen OM, Breiderhoff T, et al. Sort1, encoded by the cardiovascular risk locus 1p13.3, is a regulator of hepatic lipoprotein export. *Cell Metab*. 2010;12(3):213-223. doi:10.1016/j.cmet.2010.08.006.

93. Musunuru K, Strong A, Frank-Kamenetsky M, et al. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature*. 2010;466(7307):714-719. doi:10.1038/nature09266.

94. Grallert H, Dupuis J, Bis JC, et al. Eight genetic loci associated with variation in lipoprotein-associated phospholipase A2 mass and activity and coronary heart disease: Meta-analysis of genome-wide association studies from five community-based studies. *Eur Heart J*. 2012;33(2):238-251. doi:10.1093/eurheartj/ehr372.

95. Th?riault S, Don-Wauchope A, Chong M, Lali R, Morrison KM, Par? G. Frameshift mutation in the APOA5 gene causing hypertriglyceridemia in a Pakistani family: Management and considerations for cardiovascular risk. *J Clin Lipidol*. 2016;10(5):1272-1277. doi:10.1016/j.jacl.2016.07.009.

96. Emoto N, Yanagisawa M. Endothelin-converting enzyme-2 is a membrane-bound, phosphoramidon-sensitive metalloprotease with acidic pH optimum. *J Biol Chem*. 1995;270(25):15262-15268. doi:10.1074/jbc.270.25.15262.

97. Roselló-Lletí E, Rivera M, Miró V, Cortés R, Martínez-Dolz L, Portolés M. [Prognostic value of big endothelin-1 prognostic value in patients with heart failure and moderately symptomatic functional class]. *Med Clin (Barc)*. 2009;133(5):173-176. doi:10.1016/j.medcli.2008.10.062.

98. Zhou B-Y, Guo Y-L, Wu N-Q, et al. Plasma big endothelin-1 levels at admission and future cardiovascular outcomes: A cohort study in patients with stable coronary artery disease. *Int J Cardiol*. 2017;230:76-79. doi:http://dx.doi.org/10.1016/j.ijcard.2016.12.082.

99. Qing P, Li XL, Zhang Y, et al. Association of big endothelin-1 with coronary artery calcification. *PLoS One*. 2015;10(11). doi:10.1371/journal.pone.0142458.

100. Vacek TP, Rehman S, Neamtu D, Yu S, Givimani S, Tyagi SC. Matrix metalloproteinases in atherosclerosis: Role of nitric oxide, hydrogen sulfide, homocysteine, and polymorphisms. *Vasc Health Risk Manag*. 2015;11:173-183. doi:10.2147/VHRM.S68415.

101. Johnson JL, George SJ, Newby AC, Jackson CL. Divergent effects of matrix metalloproteinases 3, 7, 9, and 12 on atherosclerotic plaque stability in mouse brachiocephalic arteries. *Proc Natl Acad Sci*. 2005;102(43):15575-15580. doi:10.1073/pnas.0506201102.

102. Wu H, Bai X, Chen D, Cao H, Qin L. Association of Genetic Polymorphisms in Matrix Metalloproteinase-9 and Coronary Artery Disease in the Chinese Han Population: A Case–Control Study. *Genet Test Mol Biomarkers*. 2013;17(9):707-712. doi:10.1089/gtmb.2013.0109.

103. Fernandez-Patron C, Martinez-Cuesta MA, Salas E, et al. Differential regulation of platelet aggregation by matrix metalloproteinases-9 and -2. *Thromb Haemost*. 1999;82(6):1730-1735. doi:99121730 [pii].

104. Inokubo Y, Hanada H, Ishizaka H, Fukushi T, Kamada T, Okumura K. Plasma levels of matrix metalloproteinase-9 and tissue inhibitor of metalloproteinase-1 are increased in the coronary circulation in patients with acute coronary syndrome. *Am Heart J*. 2001;141(2):211-217. doi:10.1067/mhj.2001.112238.

105. Bates JM, Akerlund J, Mittge E, Guillemin K. Intestinal Alkaline Phosphatase Detoxifies Lipopolysaccharide and Prevents Inflammation in Zebrafish in

Response to the Gut Microbiota. *Cell Host Microbe*. 2007;2(6):371-382. doi:10.1016/j.chom.2007.10.010.

106. Narisawa S, Huang L, Iwasaki A, Hasegawa H, Alpers DH, Millán JL. Accelerated fat absorption in intestinal alkaline phosphatase knockout mice. *Mol Cell Biol*. 2003;23(21):7525-7530. doi:10.1128/MCB.23.21.7525.

107. Cheng JB, Jacquemin E, Gerhardt M, et al. Molecular genetics of 3beta-hydroxy-Delta5-C27-steroid oxidoreductase deficiency in 16 patients with loss of bile acid synthesis and liver disease. *J Clin Endocrinol Metab*. 2003;88(4):1833-1841. doi:10.1210/jc.2002-021580.

108. Nouvion A-L, Oubaha M, LeBlanc S, et al. CEACAM1: a key regulator of vascular permeability. *J Cell Sci*. 2010;123(24):4221-4230. doi:10.1242/jcs.073635.

109. Najjar SM, Ledford KJ, Abdallah SL, et al. Ceacam1 deletion causes vascular alterations in large vessels. *Am J Physiol Endocrinol Metab*. 2013;305(4):E519-E529. doi:10.1152/ajpendo.00266.2013.

110. Ledford KJ. Abstract 17307: Mice with Genetic Ablation of Ceacam1 Develop Spontaneous Atherosclerosis. *Circulation*. 2010. http://circ.ahajournals.org/content/122/Suppl_21/A17307.

111. Fox ER, Young JH, Li Y, et al. Association of genetic variation with systolic and diastolic blood pressure among African-Americans: The candidate gene association resource study. *Hum Mol Genet*. 2011;20(11):2273-2284. doi:ddr092 [pii]\n10.1093/hmg/ddr092.

112. Hotta K, Kitamoto T, Kitamoto A, et al. Association of variations in the FTO, SCG3 and MTMR9 genes with metabolic syndrome in a Japanese population. *J Hum Genet*. 2011;56(9):647-651. doi:10.1038/jhg.2011.74.

113. Linder P. Dead-box proteins: A family affair - Active and passive players in RNP-remodeling. *Nucleic Acids Res*. 2006;34(15):4168-4180. doi:10.1093/nar/gkl468.

114. Hug N, Cáceres JF. The RNA Helicase DHX34 Activates NMD by promoting a transition from the surveillance to the decay-inducing complex. *Cell Rep*. 2014;8(6):1845-1856. doi:10.1016/j.celrep.2014.08.020.

115. Chang Y-F, Imam JS, Wilkinson MF. The Nonsense-Mediated Decay RNA Surveillance Pathway. *Annu Rev Biochem*. 2007;76(1):51-74. doi:10.1146/annurev.biochem.76.050106.093909.

116. Yamazaki T, Goya I, Graf D, Craig S, Martin-Orozco N, Dong C. A butyrophilin family member critically inhibits T cell activation. *J Immunol*. 2010;185(10):5907-5914. doi:10.4049/jimmunol.1000835.

117. Clark MR, Mandal M, Ochiai K, Singh H. Orchestrating B cell lymphopoiesis through interplay of IL-7 receptor and pre-B cell receptor signalling. *Nat Rev Immunol*. 2013;14(2):69-80. doi:10.1038/nri3570.

118. Du?? D, Krawczenko A, Za????cki P, et al. IL-7 receptor is present on human microvascular endothelial cells. *Immunol Lett*. 2003;86(2):163-168. doi:10.1016/S0165-2478(03)00018-X.

119. Wilson GJ, Marakalala MJ, Hoving JC, et al. The C-type lectin receptor CLECSF8/CLEC4D is a key component of anti-mycobacterial immunity. *Cell Host Microbe*. 2015;17(2):252-259. doi:10.1016/j.chom.2015.01.004.

120. Wilfert AB, Chao KR, Kaushal M, et al. Genome-wide significance testing of variation from single case exomes. *Nat Genet*. 2016;48(12):1455-1461. doi:10.1038/ng.3697.

121. Zook JM, Chapman B, Wang J, et al. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat Biotechnol*. 2014;32(3):246-251. doi:10.1038/nbt.2835.

122. Coriell Cell Repositories. https://catalog.coriell.org/.

123. Real Time Genomics. https://www.realtimegenomics.com/products/rtg-tools.

124. Tassi F, Ghirotto S, Mezzavilla M, Vilaça ST, De Santi L, Barbujani G. Early modern human dispersal from Africa: genomic evidence for multiple waves of migration. *Investig Genet*. 2015;6(1):13. doi:10.1186/s13323-015-0030-2.

125. Dennell R, Roebroeks W. An Asian perspective on early human dispersal from Africa. *Nature*. 2005;438(7071):1099-1104. doi:10.1038/nature04259.

126. Nelson MR, Wegmann D, Ehm MG, et al. An Abundance of Rare Functional Variants in 202 Drug Target Genes Sequenced in 14,002 People. *Science (80-).* 2012;337(6090):100-104. doi:10.1126/science.1217876.

Supplementary Information

Table S1: Sanger sequencing primers.

DECODE ID	0036
Variant ID	rs121908031
Coding DNA change	c.2043C>A
Protien change	p.Cys681X
Forward primer	5'-ATGATCTCGTTCCTGCCCTG-3'
Reverse primer	5'-CAGAGAGAGGCTCAGGAGGG-3'
Length of amplicon	259 base pairs

DECODE ID	59
Variant ID	rs551747280
Coding DNA change	c.82G>T
Protien change	p.E28K
Forward primer	5'- AGACACAGGAAACGTGGTCA -3'
Reverse primer	5'- CTCCTGGGACTCATCAGAGC -3'
Length of amplicon	186 base pairs

DECODE ID	68
Variant ID	rs397509365
Coding DNA change	c.1690A>C
Protien change	p.Asn564His
Forward primer	5'- ACTGGACTGACTGGGGAACT -3'
Reverse primer	5'- AGCTTGGGCTTGTCCCAGA -3'
Length of amplicon	203 base pairs

DECODE ID	68
Variant ID	NA
Coding DNA change	c.2393_2402delTCCTCGTCT
Protein change	p.Leu799_Phe801del
Forward primer	5'- GGTACGATGCCCGTGTTTTC -3'
Reverse primer	5'- TGGTTGTGGCAAATGTGGAC -3'
Length of amplicon	225 base pairs

DECODE ID	20 & 42
Variant ID	rs268
Coding DNA change	c.82G>T

Protien change	p.E28K
Forward primer	Hs00784204_CE †
Reverse primer	Hs00784204_CE †
Length of amplicon	213 base pairs

† Identification of Thermofisher proprietary primers

Table S2: Coverage information for 24 WDLV genes.

Gene	**Effect of pathogenic variant**	**Syndrome**	**Mean % 20X coverage +/- SD**
LDLRAP1	Inc LDL-C	Familail Hypercholesterolemia	86.29864 +/- 5.487477
PCSK9	Dec LDL-C	Familail Hypercholesterolemia	85.48926 +/- 13.316117
ANGPTL3	Dec LDL-C; Dec TG	Hypertriglyceridemia	84.45083 +/- 14.097855
APOB	Inc/Dec LDL-C	FH	90.91872 +/- 5.077927
ABCG5	Inc sitosterol/campesterol	Sitosterolemia	81.27876 +/- 12.165487
ABCG8	Inc sitosterol/campesterol	Sitosterolemia	91.89265 +/- 9.918207
MTTP	Dec LDL-C	Hypercholesterolemia	90.84445 +/- 5.982988
SAR1B	Dec LDL-C	Hypercholesterolemia	91.99373 +/- 10.211056
LPL	Inc LDL-C; Inc TG; Dec HDL-C	FCH; FLD	86.90728 +/- 4.736173
GPIHBP1	Inc TG	Hypertriglyceridemia	60.36872 +/- 21.877923
ABCA1	Dec HDL-C	Hypoalphaproteinemia	92.99476 +/- 3.824263
LIPA	-	-	87.29760 +/- 10.902002
APOA5	Inc TG	Hypertriglyceridemia	95.68654 +/- 4.895916
APOC3	Dec TG	Hypertriglyceridemia	86.28375 +/- 25.292422
APOA1	Dec HDL-C	Hypoalphaproteinemia	85.79201 +/- 11.190626
SCARB1	-	Premature CAD	82.58899 +/- 15.299940
LIPC	Inc TG; Inc HDL-C	Hypertriglyceridemia	93.40729 +/- 7.815569
LMF1	Inc TG	Hypertriglyceridemia	89.02909 +/- 15.997057
CETP	Inc HDL-C	Hypoalphaproteinemia	94.23773 +/- 7.927210

LCAT	Dec HDL-C	Hypoalphaproteinemia	86.38898 +/- 15.349061
LIPG	Inc HDL-C	Hypoalphaproteinemia	98.14990 +/- 3.764887
LDLR	Inc LDL-C	Familail Hypercholesterolemia	90.82657 +/- 8.040695
APOE	Inc IDL-C	Familail Hypercholesterolemia	52.95406 +/- 24.454529
APOC2	Inc TG	Hypertriglyceridemia	95.48983 +/- 9.185995

Table S3: Descriptions of ACMG/AMP criteria selected as positive in InterVar to determine pathogenicity of rs551747280 observed in DECODE 59.

ACMG/AMP guideline code †	Description
PP2	Missense variant in a gene that has a low rate of benign missense variation and in which missense variants are a common mechanism of disease
PP4	Patient's phenotype or family history is highly specific for a disease with a single genetic etiology
BS2	Observed in a healthy adult individual for a recessive (homozygous), dominant (heterozygous), or X-linked (hemizygous) disorder, with full penetrance expected

† Criteria BP1 ("Missense variant in a gene for which primarily truncating variants are known to cause disease") was initially set to positive for this variant by InterVar. However, it was manually set to negative given the extensive evidence supporting causality of multiple missense variants in LDLR for FH.

Table S4: Exome sequencing datasets contributing to ExAC.

Project title
1000 Genomes Project
Bulgarian Trios
Finland-United States Investigation of NIDDM Genetics (FUSION)
GoT2D
Inflammatory Bowel Diseae
METabolic Syndrome in Men (METSIM)
Jackson Heart Study
Myocardial Infarction Genetics Consortium
NHLBI-GO Exome Sequencing Project
National Institute of Mental health (NIMH) controls

SIGMA-T2D
Sequencing in Suomi (SiSu)
Swedish Schizophrenia and Bipolar Studies
T2D-GENES
Schizophrenia Trios from Taiwan
The Cancer Genome Atlas (TCGA)
Tourrete Syndrome Association International Consortium for Genomics (TSAICG)

Table S5: Proportion of individuals from 5 community-based studies used to generate the EOMI cohort by NHLBI GO ESP6500.

Study	**Proportion in EOMI with phenotype data (N=736)**
PennCATH	36/736 (4.9%)
Cleveland Clinic Genebank	40/736 (5.4%)
Massachusetts General Hospital Premature Coronary Artery Disease Study (MGH-PCAD)	154/736 (20.9%)
Heart Attack Risk in Puget Sound (HARPS)	428/736 (58.1%)
Transnational Research Investigation Underlying Disparities in Myocardial Infarction Patients' Health Studies (TRIUMPH)	78/736 (10.6%)

Table S6: Exome-wide significant cutoffs for different combinations of association model, MAF threshold and pathogenicity criteria in EOMI Europeans and Africans using the additive model of inheritance.

	T1 alleles EOMI EUR	T5 alleles EOMI EUR	T1 alleles EOMI AFR	T5 alleles EOMI AFR
All nonsynonymous SNV + disruptive	1.655e-05	1.47e-05	2.52e-05	1.845e-05
Nonsynonymous SNV predicted deleterious or damaging by SIFT or PP2-HDIV/HVAR + disruptive	1.83e-05	1.67e-05	2.76e-05	2.12e-05
Nonsynonymous SNV with CADD score > 20 + disruptive variants	1.96e-05	1.805e-05	2.925e-05	2.29e-05

Nonsynonymous SNV predicted to be damaging by M-CAP + disruptive variants	2.06e-05	2.035e-05	3.28e-05	2.8e-05

Table S7: Exome-wide significant cutoffs for different combinations of association model, MAF threshold and pathogenicity criteria in DECODE

	T1 alleles DECODE	T5 alleles DECODE
All nonsynonymous SNV + disruptive	Add: 3.78e-05	Add: 3.615e-05 Rec: 1.41e-04
Nonsynonymous SNV predicted deleterious or damaging by SIFT or PP2-HDIV/HVAR + disruptive	Add: 5.355e-05	Add: 4.74e-05 Rec: 3.17e-03
Nonsynonymous SNV with CADD score > 20 + disruptive variants	Add: 5.91e-05	Add: 5.305e-05 Rec: 6.64e-03
Nonsynonymous SNV predicted to be damaging by M-CAP + disruptive variants	Add: 7.34e-05	Add: 7.125e-05 Rec:8.84e-03

Add = additive; rec = recessive

Table S8: 71 genes coding for proteins encompassing the endothelial secretome. Obtained from Tunica *et al.* 2009

Gene	**Category**
LAMP1	Lysosomal
CTSD	Lysosomal
PRCP	Lysosomal
SAP	Lysosomal
P4HB	Miscellaneous membrane proteins
PDIA3	Miscellaneous membrane proteins
VAS1	Miscellaneous membrane proteins
LMAN2	Miscellaneous membrane proteins
VAT1	Miscellaneous membrane proteins
THBS1	Membrane antigens and receptors

MCAM	Membrane antigens and receptors
CDH5	Membrane antigens and receptors
AXL	Membrane antigens and receptors
PROCR	Membrane antigens and receptors
CD93	Membrane antigens and receptors
ICAM2	Membrane antigens and receptors
CD59	Membrane antigens and receptors
MANF	Miscellaneous secreted proteins
B2M	Miscellaneous secreted proteins
CLSTN1	Annexins and calcium ion-binding proteins
CALR	Annexins and calcium ion-binding proteins
CALU	Annexins and calcium ion-binding proteins
ANXA5	Annexins and calcium ion-binding proteins
ANXA2	Annexins and calcium ion-binding proteins
MMRN1	Coagulation and related proteins
TFPI	Coagulation and related proteins
SERPINE1	Coagulation and related proteins
VWF	Coagulation and related proteins
S100A9	Protein S100 family
S100A8	Protein S100 family
S100A7	Protein S100 family
CFI	Inflammation-related proteins
C4A	Inflammation-related proteins
C4B	Inflammation-related proteins
APP	Inflammation-related proteins
MYDG5	Inflammation-related proteins
PTX3	Inflammation-related proteins
IL1RL1	Inflammation-related proteins
CXCL1	Growth factors and related proteins
DKK3	Growth factors and related proteins
FSTL1	Growth factors and related proteins
CTGF	Growth factors and related proteins
CYR61	Insulin-like growth factor-binding proteins

IGFBP7	Insulin-like growth factor-binding proteins
IGFBP4	Insulin-like growth factor-binding proteins
IGFBP2	Insulin-like growth factor-binding proteins
MMP1	Proteinases
MMP2	Proteinases
MMP14	Proteinases
ADAM9	Proteinases
ADAM10	Proteinases
ADAM15	Proteinases
ANPEP	Proteinases
PRSS23	Proteinases
TIMP1	Proteinase inhibitors
CST3	Proteinase inhibitors
ITIH2	Proteinase inhibitors
A2M	Proteinase inhibitors
LOXL2	Extracellular matrix components
TGM2	Extracellular matrix components
COL5A2	Extracellular matrix components
COL4A2	Extracellular matrix components
EMILIN3	Extracellular matrix components
MMRN2	Extracellular matrix components
SPOCK1	Extracellular matrix components
EFEMP1	Extracellular matrix components
FBN1	Extracellular matrix components
LGALS3	Extracellular matrix components
LGALS1	Extracellular matrix components
HSPG2	Extracellular matrix components
SPARC	Extracellular matrix components

Table S9: Mutation loads for ExAC and local samples along with correction factors for 39 DECODE European and 77 CAD-free samples

ID	ExAC mutation load	sample mutation load	CF (sample mutation load/ ExAC mutation load)	cohort
0004	64.461	99	1.535812352	DECODE
0006	83.2521	88	1.057030393	DECODE
0008	82.7825	86	1.03886691	DECODE
0009	81.3824	91	1.118177886	DECODE
0010	81.4732	85	1.043287854	DECODE
0011	80.2178	89	1.109479442	DECODE
0012	72.4228	84	1.159855736	DECODE
0015	64.8238	79	1.218688198	DECODE
0016	64.7258	61	0.942437173	DECODE
0017	82.2134	90	1.094712054	DECODE
0019	69.3246	89	1.283815558	DECODE
0033	75.8458	72	0.94929449	DECODE
0034	80.5963	82	1.017416432	DECODE
0035	80.2764	89	1.108669547	DECODE
1017	75.1895	64	0.851182678	DECODE
20	77.5955	66	0.850564788	DECODE
31	75.8624	71	0.935905007	DECODE
37	82.9627	73	0.879913503	DECODE
38	71.2326	74	1.038850189	DECODE
39	76.9544	86	1.117544936	DECODE
40	71.0045	78	1.098521925	DECODE
41	76.6651	70	0.913062136	DECODE
42	59.1335	48	0.811722628	DECODE
43	64.0268	57	0.890252207	DECODE
44	67.4057	51	0.756612571	DECODE
45	65.5763	58	0.884465882	DECODE
47	79.6407	98	1.230526603	DECODE

48	78.0061	80	1.025560822	DECODE
55	77.0238	79	1.025657005	DECODE
56	76.4724	76	0.993822608	DECODE
57	76.8461	86	1.119119903	DECODE
59	79.2117	94	1.18669338	DECODE
60	78.9517	68	0.861286077	DECODE
61	75.1944	66	0.877724937	DECODE
64	76.2818	78	1.022524377	DECODE
65	77.0448	81	1.051336365	DECODE
66	79.4132	75	0.944427375	DECODE
67	61.8038	63	1.019354797	DECODE
68	79.1885	94	1.187041048	DECODE
1021L	67.048	70	1.044028159	CAD-free
1030L	74.7556	70	0.936384699	CAD-free
1032L	73.6987	73	0.990519507	CAD-free
1033L	69.8364	75	1.073938519	CAD-free
1034L	73.6308	82	1.11366439	CAD-free
1043L	73.0803	65	0.889432583	CAD-free
1045L	74.257	68	0.915738584	CAD-free
1069L	66.9176	77	1.150668882	CAD-free
1072L	70.534	82	1.1625599	CAD-free
1074L	64.5104	102	1.581140405	CAD-free
1076L	79.3039	82	1.033997067	CAD-free
1092L	73.2635	73	0.996403393	CAD-free
1110L	73.9843	67	0.905597539	CAD-free
1112L	73.3934	70	0.953764235	CAD-free
1114L	71.8858	62	0.862479099	CAD-free
1115L	75.5051	75	0.993310386	CAD-free
1118L	76.2305	80	1.049448711	CAD-free
1124L	76.6971	76	0.990910999	CAD-free
1125L	78.5164	78	0.99342303	CAD-free
250	64.9864	67	1.030984945	CAD-free
252	60.6745	58	0.955920527	CAD-free
376	67.4408	65	0.963808259	CAD-free

384	76.4858	85	1.111317395	CAD-free
484	75.244	76	1.010047313	CAD-free
487	69.1749	79	1.142032732	CAD-free
514	76.7069	92	1.199370591	CAD-free
515	76.2442	85	1.114838899	CAD-free
552	76.3149	86	1.126909686	CAD-free
567	76.5789	82	1.07079104	CAD-free
572	77.5614	86	1.108798964	CAD-free
573	66.7109	82	1.229184436	CAD-free
593	77.9985	71	0.910273916	CAD-free
608	77.554	75	0.967068107	CAD-free
621	71.3016	82	1.150044319	CAD-free
624	73.7562	86	1.166003672	CAD-free
641	72.9285	80	1.096964835	CAD-free
645	75.1716	77	1.024323016	CAD-free
656	76.5532	75	0.979710842	CAD-free
671	63.0993	56	0.887490036	CAD-free
690	71.1258	68	0.956052515	CAD-free
695	76.7941	64	0.833397357	CAD-free
696	72.1456	84	1.164312169	CAD-free
699	68.7925	96	1.395500963	CAD-free
704	31.4668	33	1.04872437	CAD-free
708	77.2935	71	0.918576594	CAD-free
723	73.8917	70	0.94733238	CAD-free
724	67.0486	68	1.014189707	CAD-free
729	29.9163	44	1.470770115	CAD-free
780	53.5542	63	1.176378323	CAD-free
782	77.3321	75	0.969843054	CAD-free
787	66.5684	61	0.916350701	CAD-free
790	67.8881	68	1.001648301	CAD-free
811	69.3525	83	1.196784543	CAD-free
817	78.0461	74	0.948157563	CAD-free
819	72.2124	62	0.858578305	CAD-free
820	67.5497	76	1.125097521	CAD-free

830	70.1208	60	0.855666222	CAD-free
833	72.4961	70	0.965569182	CAD-free
860	72.4973	80	1.103489371	CAD-free
863	75.3034	74	0.982691353	CAD-free
866	73.5939	74	1.00551812	CAD-free
871	73.3058	76	1.03675289	CAD-free
930	74.5059	77	1.033475201	CAD-free
943	21.6206	14	0.647530596	CAD-free
947	51.9509	62	1.19343457	CAD-free
958	77.8093	68	0.873931522	CAD-free
962	75.7655	88	1.161478509	CAD-free
967	69.7341	78	1.118534548	CAD-free
975	67.6827	77	1.13766147	CAD-free
977	74.6568	73	0.977807782	CAD-free
979	76.6495	77	1.004572763	CAD-free
988	70.8032	56	0.790924704	CAD-free
997	74.7549	67	0.896262319	CAD-free
999	73.7487	77	1.044086201	CAD-free
B574	74.2564	57	0.767610603	CAD-free
B633	66.6077	85	1.276128736	CAD-free
C576	76.2868	62	0.812722516	CAD-free

Table S10: CAD ORs and z-scores before and after incorporation of the CF in DECODE Europeans.

ID	**CAD OR**	**CF-adjusted CAD OR**	**z-score**	**CF-adjusted z-score**
0004	2.739918065	1.561207353	1.322725573	0.58458602
0006	1.740582002	1.635116702	0.727314201	0.645287073
0008	5.547526686	5.086733771	2.248468854	2.134669249
0009	0.546459943	0.546459943	-0.793028345	-0.793028345
0010	0.546459943	0.546459943	-0.793028345	-0.793028345
0011	2.894506461	2.455448788	1.394754437	1.178870907
0012	0.830191431	0.783693279	-0.244221667	-0.319862033

0015	0.546459943	0.546459943	-0.793028345	-0.793028345
0016	7.45848602	8.749452014	2.636916351	2.846413544
0017	2.172766682	1.928193555	1.018363196	0.86164873
0019	1.660495524	1.29876195	0.6654993	0.343055876
0033	2.145594915	2.308210103	1.001848338	1.097720657
0034	16.37829462	15.45200462	3.669194338	3.592793324
0035	0.546459943	0.546459943	-0.793028345	-0.793028345
1017	16.72799952	30.42492826	3.69691972	4.481922006
20	2.631421308	3.468337878	1.269702674	1.632102032
31	0.546459943	0.546459943	-0.793028345	-0.793028345
37	1.484395558	1.701290515	0.518377029	0.697350685
38	0.546459943	0.546459943	-0.793028345	-0.793028345
39	0.546459943	0.546459943	-0.793028345	-0.793028345
40	0.546459943	0.546459943	-0.793028345	-0.793028345
41	0.546459943	0.546459943	-0.793028345	-0.793028345
42	0.806315542	0.882453378	-0.282516726	-0.164104908
43	0.546459943	0.546459943	-0.793028345	-0.793028345
44	0.546459943	0.546459943	-0.793028345	-0.793028345
45	0.546459943	0.546459943	-0.793028345	-0.793028345
47	0.546459943	0.546459943	-0.793028345	-0.793028345
48	0.546459943	0.546459943	-0.793028345	-0.793028345
55	2.854427417	2.738779304	1.376456282	1.322180035
56	0.546459943	0.546459943	-0.793028345	-0.793028345
57	0.546459943	0.546459943	-0.793028345	-0.793028345
59	0.546459943	0.546459943	-0.793028345	-0.793028345
60	0.546459943	0.546459943	-0.793028345	-0.793028345
61	3.735659512	4.88274106	1.729540455	2.080957014
64	1.42636625	1.396542788	0.466044888	0.438315036
65	0.546459943	0.546459943	-0.793028345	-0.793028345
66 †	**3.29605432**	**3.663680773**	**1.565240055**	**1.704007847**
67	2.556323862	2.482540732	1.231705868	1.193270976
68	1.314701978	1.144857333	0.359064286	0.177530887

† Significant z-score after CF adjustment