6-7-2013 12:00 AM

Genetic approaches to studying complex human disease

Joseph B. Dube, *The University of Western Ontario*

Supervisor: Dr. Robert A. Hegele, *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Master of Science degree in Biochemistry

Genetic approaches to studying complex human disease

(Thesis format: monograph)

by

Joseph Brenton Dubé

Graduate program in Biochemistry

A thesis submitted in partial fulfillment of
the requirements for the degree of
Master of Science

The School of Graduate and Postdoctoral Studies
The University of Western Ontario
London, Ontario, Canada

ABSTRACT

Common, complex diseases such as cardiovascular disease (CVD) represent an intricate interaction between environmental and genetic factors and now account for the leading causes of mortality in western society. By investigating the genetic component of complex disease etiology, we have gained a better understanding of the biological pathways underlying complex disease and the heterogeneity of complex disease risk. However, the development of high throughput genomic technologies and large well-phenotyped multi-ethnic cohorts has opened the door towards more in-depth and trans-disciplinary approaches to studying the genetics of complex disease pathogenesis. Accordingly, we sought to investigate select complex traits and diseases using both established and novel genomic technologies, including candidate gene resequencing, high-throughput targeted microarray genotyping and candidate variant genotyping. We demonstrate that a private and common variant, p.G116S, within the low-density lipoprotein receptor (*LDLR*) gene among Inuit descendants has a large effect on plasma cholesterol; that variation in cardio-metabolic and Alzheimer disease (AD) loci is not associated with susceptibility to the pre-dementia phenotype known as "cognitive impairment, no dementia"; and that established type 2 diabetes (T2D) variants are not associated with T2D susceptibility among select aboriginal Canadian and Greenland cohorts. Together, these studies represent a selection of established and novel genomic strategies for the investigation of complex disease genetics which are likely to remain fundamental in the continued investigation of complex disease pathogenesis.

KEYWORDS: cardiovascular disease, dementia, Alzheimer disease, vascular dementia, "cognitive impairment, no dementia", mild cognitive impairment, type 2 diabetes, hypercholesterolemia, low-density lipoprotein cholesterol, apolipoprotein E, complex disease, genetic variation, single nucleotide polymorphisms, genome-wide association studies, genetic risk scores.

CO-AUTHORSHIP

References for published material in this dissertation are listed at the beginning of each respective chapter. This section describes the contributions of co-authors.

Dr. Robert A. Hegele (supervisor) provided funding, supervision, and samples for all studies. He also contributed to study design, manuscript preparation and critical revision for all chapters.

Drs. T. Kue Young, Eric Dewailly, Peter Bjerregaard, and Bert B. Boyer provided Inuit population-based samples and clinical data used in chapter 2. Randa Stringer provided genotyping used in chapter 2.

Dr. Jian Wang and Dr. Henian Cao managed clinical databases and provided excellent technical assistance and supervision for genotyping performed in chapters 2-4.

Dr. Christopher T. Johansen provided supervision, contributed to statistical analyses and provided critical review comments for the manuscript in chapter 3. Drs. Vladimir Hachinski and Joan Lindsay provided samples and clinical data from the Canadian Study of Health and Aging (CSHA) used in chapter 3 and also provided critical review comments for the manuscript in chapter 3.

Adam D. McIntyre, John Robinson, and Matthew R. Ban provided technical assistance for chapters 2-4. Sean J. Leith performed genotyping used in chapter 4.

DEDICATION

For my family, and my mentor Dr. Hegele.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

LIST OF ABBREVIATIONS

3MS	**Modified Mini-Mental State Exam**
ABCA7	**ATP-binding cassette, sub-family A (ABC1), member 7**
ABCG2	**ATP-binding cassette, sub-family G (WHITE), member 2**
AD	**Alzheimer disease**
ADAMTS9	**ADAM metallopeptidase with thrombospondin type 1 motif, 9**
ADR	**adverse drug reaction**
AP3S2	**adaptor-related protein complex 3, sigma 2 subunit**
APOB	**apolipoprotein B**
APOE	**apolipoprotein E**
APP	**amyloid precursor protein**
ASO	**anti-sense oligonucleotide**
Aβ	**β-amyloid**
BLAT	**Basic Local Alignment Search Tool**
BMI	**body mass index**
CADASIL	**cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy**
CDCV	**common-disease common-variant**
CDKAL1	**CDK5 regulatory subunit associated protein 1-like 1**
CDKN2A/2B	**cyclin-dependent kinase inhibitor 2A/2B**
CHD	**coronary heart disease**
CHR	**chromosome**
C-IMT	**common carotid intima-media thickness**
CIND	**"cognitive impairment, no dementia"**

CPT1A	**carnitine palmitoyltransferase**
CR1	**complement component (3b/4b) receptor 1**
CSHA	**Canadian Study of Health and Aging**
CVD	**cardiovascular disease**
diLQTS	**drug-induced long QT syndrome**
DSM	**Diagnostic and Statistical Manual of Mental Disorders**
FBG	**fasting blood glucose**
FH	**familial hypercholesterolemia**
GCKR	**glucokinase (hexokinase 4) regulator**
GPIHBP1	**glycosylphosphatidylinositol anchored high density lipoprotein binding protein 1**
GRAMD3	**GRAM domain containing 3**
GRS	**genetic risk score**
GWAS	**genome-wide association study**
HDL	**high-density lipoprotein**
HDL-C	**HDL cholesterol**
HeFH	**heterozygous FH**
HGMD	**Human Gene Mutation Database**
HIV	**human immunodeficiency virus**
HMG20A	**high mobility group 20A**
HMGCR	**3-hydroxy-3-methylglutaryl-CoA reductase**
HNF1A	**hepatocyte nuclear factor 1 homeobox A**
HNF4A	**hepatocyte nuclear factor 4, alpha**
HoFH	**homozygous FH**

HTG	**hypertriglyceridemia**
HWE	**Hardy-Weinberg Equilibrium**
IGT	**impairt glucose tolerance**
IMT	**intima-media thickness**
KCNJ11	**potassium inwardly-rectifying channel, subfamily J, member 11**
LD	**linkage disequilibrium**
LDL	**low-density lipoprotein**
LDL-C	**LDL cholesterol**
LDLR	**LDL receptor**
LMF1	**lipase maturation factor 1**
LPA	**lipoprotein, Lp(a)**
LPL	**lipoprotein lipase**
LPLD	**LPL deficiency**
MAF	**minor allele frequency**
MCI	**mild cognitive impairment**
MNTR1B	**melatonin receptor 1B**
MTTP	**microsomal triglyceride transfer protein**
MutPred	**Mutation Prediction**
NFT	**neurofibrillary tangle**
NGS	**next-generation sequencing**
NHLBI	**National Heart, Lung and Blood Institute**
OMIM	**Online Mendelian Inheritance in Man**
OR	**odds ratio**

PCSK9	**proprotein convertase subtilisin/kexin type 9**
PICALM	**phosphatidylinositol binding clathrin assembly protein**
Polyphen	**Polymorphism Phenotyping**
PPARG	**peroxisome proliferator-activated receptor gamma**
PSEN1	**presenilin 1**
PSEN2	**presenilin 2**
Q-Q	**quantile-quantile**
RPS26A	**ribosomal 40S subunit protein S26A**
SIFT	**sorting intolerant from tolerant**
SLCO1B1	**solute carrier organic anion transporter family, member 1B1**
SNP	**single nucleotide polymorphism**
SNV	**single nucleotide variant**
T2D	**type 2 diabetes**
TC	**total cholesterol**
TCF7L2	**transcription factor 7-like 2 (T-cell specific, HMG-box)**
TG	**triglyceride**
TSPAN6	**tetraspanin 6**
TSPAN8	**tetraspanin 8**
UBE2E2	**ubiquitin-conjugating enzyme E2E 2**
VaD	**vascular dementia**
ZHX2	**zinc fingers and homeoboxes 2**
ZNF608	**zinc finger protein 608**

CHAPTER 1

INTRODUCTION

This chapter is based on material from the following publications: (1) **Dube, J.B.**, and Hegele, R.A. (2012). Genetics 100 for cardiologists: basics of genome-wide association studies. *Can J Cardiol 29*, 10-17; and (2) **Dube, J.B.**, Johansen, C.T., and Hegele, R.A. (2011). Sortilin: an unusual suspect in cholesterol metabolism: from GWAS identification to in vivo biochemical analyses, sortilin has been identified as a novel mediator of human lipoprotein metabolism. *Bioessays 33*, 430-437.

1.1 Human disease

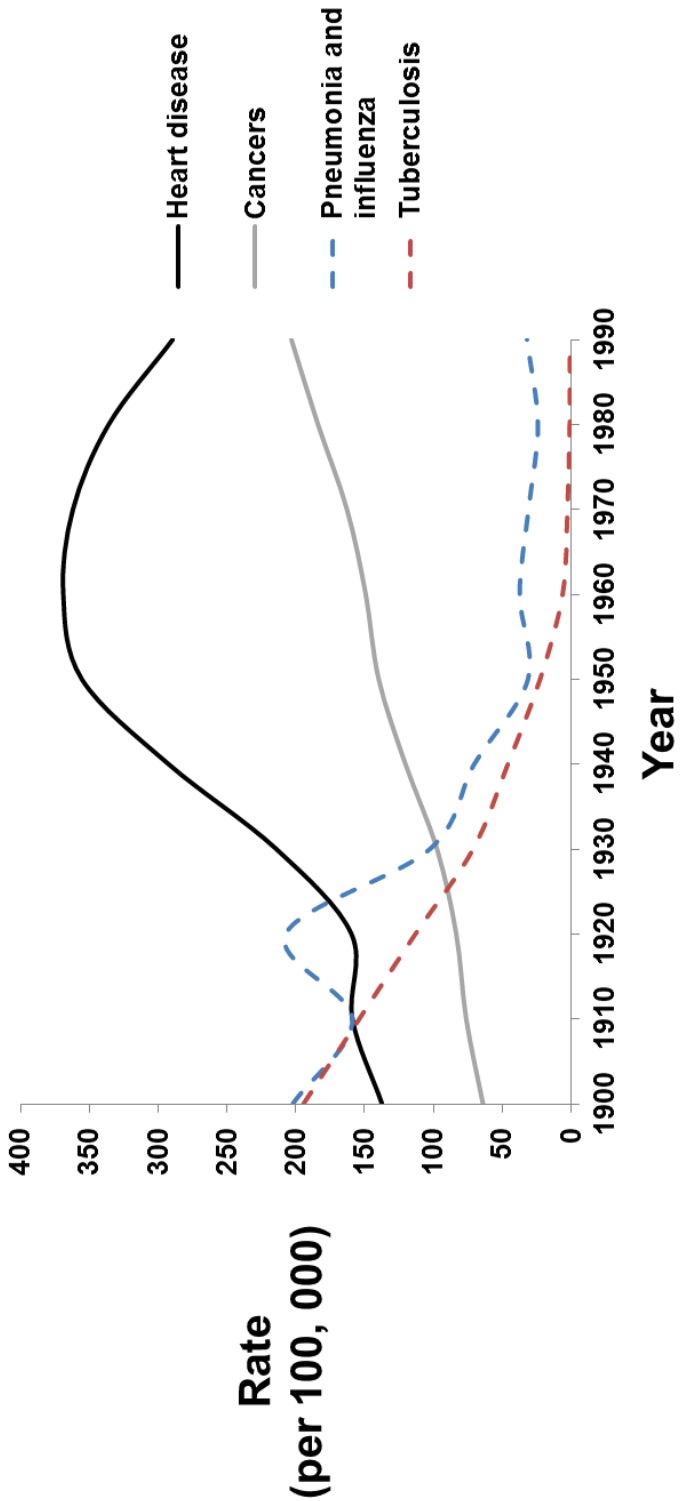
The concept of human disease can be briefly defined as dysfunction or abnormality in biological function which enhances mortality and morbidity risk. As medical science has advanced, preventive strategies have largely mitigated many of the risk factors underlying the once common infectious diseases in developed nations. Although historically notorious diseases such as the bubonic plague are no longer a leading cause of death, a new class of diseases has risen to represent the top causes of mortality in the developed world. Chronic non-infectious diseases that become clinically recognizable in adulthood have reached prevalence worldwide as in correlation with longer average lifespan. Diseases such as cancers, heart disease and stroke now account for a major

percentage of deaths worldwide and thus pose the greatest threat to modern global public health (Lozano et al., 2013).

1.1.1 History of human disease

In recent history, the most prominent diseases were infectious in nature and were often linked to hygienic deficiencies. The plague of the 14th century, one of the world's greatest epidemics, was caused by the insidious spread of *Yersinia pestis* by fleas and rats (Ligon, 2006). Tuberculosis, another bacterial disease which reached epidemic proportions in Europe during the Industrial Revolution, is believed to have been largely transmitted through unpasteurized milk and milk products (Donoghue, 2009). Discoveries in antibiotics, improved hygienic and sanitary practices as well as technological advancements have since helped to limit epidemics of infectious disease. As these advances have dramatically decreased early life mortality and have supported a longer average lifespan among Western countries, the leading causes of mortality are now represented by chronic age-related diseases (**Figure 1.1**). In North America, cancer, heart disease and stroke cumulatively account for almost half of all reported deaths (Heron, 2012). Dementia and cognitive impairment, both of which are highly correlated with age, are anticipated as the next epidemics to emerge over the coming decades. As these age-related diseases represent major public health concerns, understanding the etiologies and risk factors underlying common diseases has become a global imperative.

Figure 1.1 Historic rates of leading causes of death in the United States from 1900-1998. Data presented here were taken from the U.S. National Center for Health Statistics (Centers for Disease Control and Prevention, 2009).


Heart disease
Cancers
Pneumonia and influenza
Tuberculosis
Rate (per 100, 000)
400
350
300
250
200
150
100
50
0
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
Year

1.1.2 Heritability of disease

Diseases may be distinguished based on the mode of disease transmission. Fundamental classifications have characterized diseases as communicable and non-communicable based on the respective presence or absence of a pathogenic microorganism necessary for disease transmission. However, a crucial distinction in disease classification was the observed inheritance of diseases or traits in offspring following mathematical ratios as per Mendel's early studies in peas. Pedigrees charting the inheritance of traits within a family tree helped conceptualize a novel means for the transmission of disease susceptibility via heritable or genetic factors. The heritability of a wide range of traits and diseases has been explored through family-based studies that looked at phenotypic heritability between related individuals who consequently were less genetically heterogeneous compared to the general population. Studies comparing large sets of twin pairs were also important as concordance rates between monozygotic twins who are genetically identical could also be used in estimating the heritability of diseases. More recently, genetic studies have expanded to large cohorts representative of the general population where genetic heterogeneity is greatest. Intriguingly, family and twin studies have ascribed strong heritability estimates to the cardio-metabolic traits relating to heart disease as well as cognitive disorders such as Alzheimer disease (AD) (Mangino and Spector, 2012).

1.1.2.1 Monogenic disease

In rare cases, diseases have been observed to segregate predictably and according to Mendelian ratios among offspring. Through early genetic mapping approaches, which are

discussed later, it was shown that a single genetic variant of deleterious effect was sufficient to cause remarkably penetrant and pathogenic phenotypes. Furthermore, the variants underlying Mendelian diseases tended to disrupt the function of a single locus or gene which led to the concept of monogenic diseases. Regarding specific modes of inheritance, monogenic diseases can be inherited in an autosomal recessive manner in which two dysfunctional alleles must be inherited for the disease phenotype to manifest. Alternatively, monogenic diseases can be inherited in an autosomal dominant or co-dominant manner whereby a single dysfunctional allele can cause the disease phenotype or two dysfunctional alleles can create an even more severe phenotype. Additional modes of inheritance include sex-linked patterns of heritability. As monogenic diseases exemplify the biological effects as a result of disrupting individual genes, this class of genetic disease has been invaluable in helping improve our understanding of the genetic architecture underlying many diseases and clinically-important traits currently documented in the Online Mendelian Inheritance in Man database (OMIM) (Hamosh et al., 2005).

1.1.2.2 Complex disease

In contrast to monogenic diseases, where a single variant is sufficient to cause a disease phenotype, complex diseases represent a greater interaction between environmental and genetic factors. No single genetic variant is sufficient to cause a complex disease phenotype but rather multiple variants of low penetrance at multiple loci contribute synergistically to modulate disease susceptibility. The leading model for complex diseases is described as the common disease-common variant (CDCV) hypothesis which

predicts that multiple commonly occurring variants from multiple genes individually contribute a small effect on disease susceptibility but additively exert a considerable effect in the manifestation of complex disease (Reich and Lander, 2001). An emerging hypothesis has incorporated the potential contribution of rare variants of proportionately larger effect on complex disease susceptibility; however, this concept is currently being assessed for validity in the most common complex diseases (Pritchard, 2001). In support of the heterogeneous nature of complex disease susceptibility, complex diseases are not inherited according to the models which apply to monogenic diseases. Instead, common variants are believed to contribute to the overall picture of disease susceptibility.

1.1.2.3 Disease penetrance

The phenotypic spectrum that exists within complex polygenic diseases, in terms of characteristics such as disease severity and age of onset, is mediated by the complex interaction between environmental and genetic risk factors. For certain complex diseases, disease susceptibility is greatly modulated by variants with highly penetrant effects such as hereditary forms of breast and ovarian cancer linked with mutations in the genes encoding breast cancer 1 (*BRCA1*) and 2 (*BRCA2*) (Apostolou and Fostira, 2013). While these large-effect variants are not deterministic of disease, it is important to note that single variants can mediate patterns of inheritance similar to monogenic diseases.

1.2 Genetic variation

The human and chimpanzee genomes differ by only ~4% which is remarkable considering apparent physical distinctions (Varki and Altheide, 2005). In a more focused comparison, any two humans share ~99.5% genomic similarity and yet people vary considerably in terms of characteristics such as anthropometric measurements, eye or hair colour (Tishkoff and Kidd, 2004). Genetic variation represents the small percentage of genomic divergence and substantially contributes to the range of anthropometric traits we observe in human populations. The approaches currently used to study genetic variation are the result of the Human Genome Project which provided the first draft sequence of the human genome as well as the efforts of international consortia such as the International HapMap Project and the 1000 Genomes Project that developed a detailed map of genetic variation throughout the human genome (2003). As one of the most crucial tools to emerge from the genomic era, the comprehensive catalogue of genetic variation in humans has facilitated unprecedented analyses of the role of genetic variation in modulating various phenotypes and disease susceptibilities with novel applications continuously emerging.

1.2.1 Classes of genetic variation

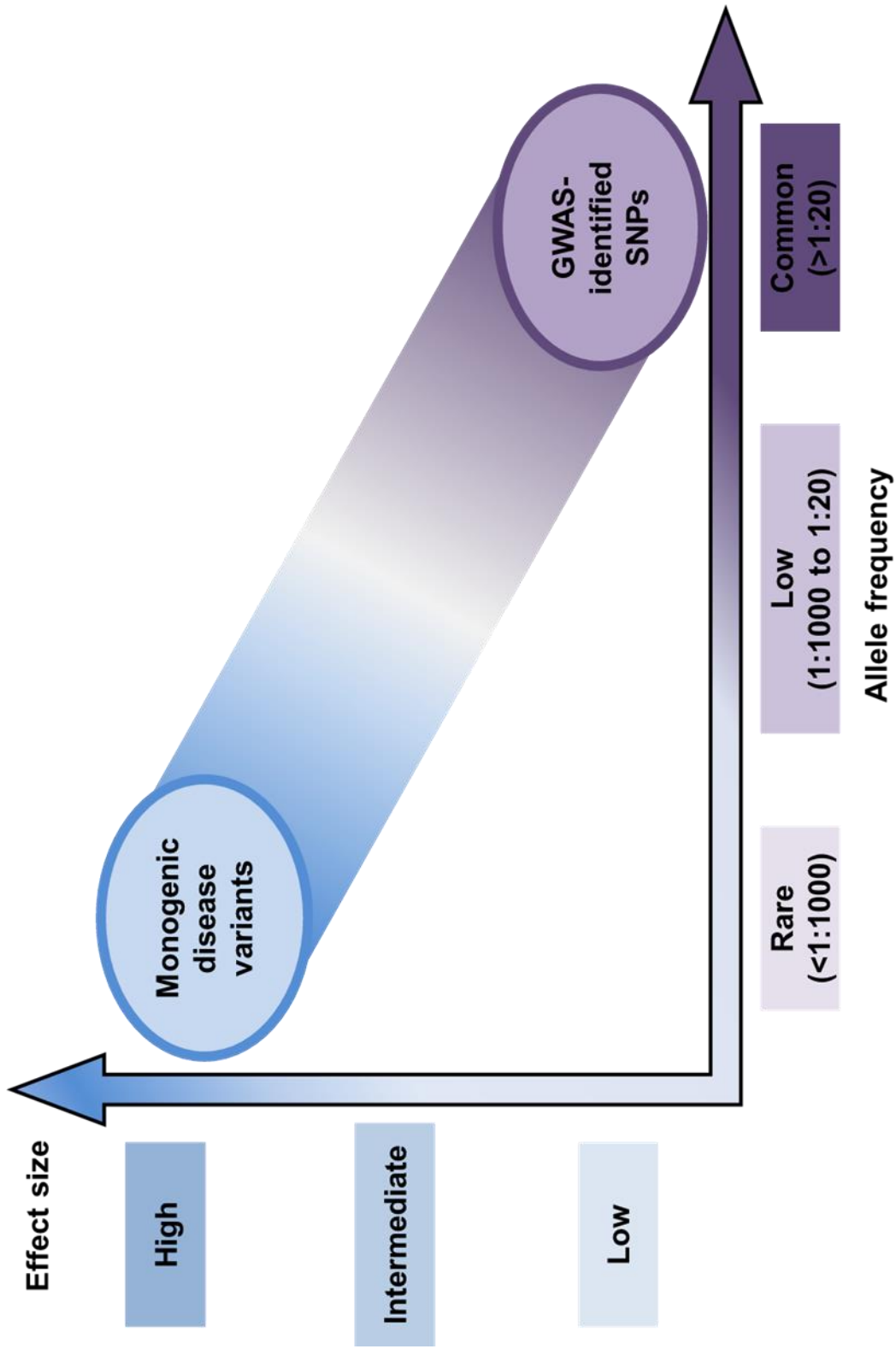
Several types of genetic variants comprise the ~0.5% of inter-individual genetic divergence and are classed in terms of size and frequency. The various classes of observed genetic variation are well documented and may vary considerably from large-scale variation in the number of copies of entire chromosomes such as trisomy of

chromosome 21 in Down syndrome to smaller-scale variation in the case of single base substitutions or single nucleotide variants (SNVs). SNVs represent the most abundant form of genetic variation and may potentially affect gene expression or protein structure based on localization. SNVs are distinguished based on the frequency with which they are observed in a population. Rare SNVs are observed with <1% frequency, uncommon SNVs occur with 1%-5% frequency and common SNVs are observed with frequencies >5%. Common SNVs, or single nucleotide polymorphisms (SNPs), have become established as important markers in genomic mapping strategies based on the global prevalence of these variants as well as the genome-wide coverage that the >38 million validated SNPs offers.

1.2.2 Effect-frequency relationship

Based on the genetic models of monogenic and complex diseases, a distinct trend has been observed between variant frequency and the variant-associated effect on disease risk (**Figure 1.2**). The frequency of causative variants underlying monogenic diseases tends to be extremely rare as these variants are subjected to heavy selective pressure and are thus not likely to be propagated in subsequent generations. At the opposite end of the frequency spectrum, SNPs have been subjected to low selective pressure due presumably to small phenotypic effects and have thus reached relatively high frequencies in global populations. The implications of this frequency-effect trend strongly influenced our concept of rare and common disease, which led to the development of the CDCV hypothesis. As common variation has been widely hypothesized to account considerably

Figure 1.2 Inverse relationship between variant frequency and variant effect size. Studies on rare monogenic disorders and common, complex diseases have supported the correlation between variant frequency and effect size whereby rare variants tend to associate with large and often deleterious effects while common variants are more likely to have subtle effects on disease susceptibility. Modified from Manolio *et al.* (2009).


Effect size
High
Intermediate
Low
Monogenic disease variants
GWAS-identified SNPs
Rare (<1:1000)
Low (1:1000 to 1:20)
Common (>1:20)
Allele frequency

for the prevalence of common and complex disease, SNPs have been established as the genomic marker of choice for the majority of studies emerging from the genomics era.

1.2.3 Linkage disequilibrium

The concept of independent assortment was a fundamental assumption in Mendel's studies of peas in which the heritable factors underlying traits were passed on to offspring independently of each other. This concept is applicable for variants separated by considerable distance or located on different chromosomes; however, in the case of SNPs that are present every ~300 bases, SNP alleles spanning a limited physical range tend to be inherited together as clusters known as haplotype blocks through a phenomenon called linkage disequilibrium (LD) (Gabriel et al., 2002). Through the genotyping of millions of SNPs in multi-ethnic populations, the International HapMap Project created a comprehensive map of haplotype block structures across the human genome as well asestimates of LD between variants, which have permitted the reliable prediction or imputation of variant genotypes. By imputing SNP genotypes, it is possible to vastly expand genomic coverage while sparing the costs of directly genotyping potentially millions of supplemental markers.

1.2.4 Hardy-Weinberg equilibrium

Another important characteristic of variant frequencies involves the allelic distribution for a given variant. SNPs are most often bi-allelic which means for any SNP there exists a major allele that represents >50% of all alleles specific to that SNP in a given population while the minor allele represents the remaining percentage. As part of the

CDCV hypothesis, SNPs are believed to individually contribute small effects towards disease susceptibility. Accordingly, the relative frequencies of homozygotes and heterozygotes for a given SNP genotype should remain stable from generation to generation. The Hardy-Weinberg equilibrium (HWE) principle is represented in a mathematical equation for calculating expected genotype frequencies for a biallelic variant, assuming the absence of such influences on allele frequency including genetic drift, natural selection or non-random mating (Mayo, 2008). Thus comparability between observed and expected genotype frequencies using HWE tests a fundamental assumption of the CDCV hypothesis. More recently, large-scale genotyping studies have used HWE as a means of statistical quality control in detecting potential genotyping errors by the variants that deviate substantially from HWE genotype proportions (Pearson and Manolio, 2008).

1.3 Approaches to studying genetic disease

The various approaches to studying the genetics of complex diseases have evolved in parallel with the development of genomic technologies. In the pre-genomic era, correlations were investigated between patients and clinically important variables such as weight, lipid profile or blood type. While these clinical characteristics served in part as surrogates for genetic variants, the first studies to directly investigate genetic diseases were based on families affected by predictably segregating disease phenotypes. The progression to population genetics studies came with a host of novel technological and statistical approaches aimed at testing variant-disease association in large genetically

diverse cohorts. Interestingly, as new technologies have been integrated in the field of medical genetics, the strategies of family- and population-based studies have remained fundamental to the ongoing study of genetic susceptibility to complex disease.

1.3.1 Family-based techniques

As the traditional approach to studying disease genetics, family-based studies have been important in identifying the genetic architecture of highly penetrant and monogenic phenotypes. Linkage analysis represents the established approach for assessing the cosegregation of trait loci specifically within families (Dawn Teare and Barrett, 2005). Methodologically, linkage analyses are best suited to the study of monogenic diseases in which a highly penetrant mutation underlies disease susceptibility. Statistical power is also derived from the exclusive study of family members where genetic variation between individuals will be minimal thus lowering the false-positive discovery rate or type 1 error. As technology advanced, linkage analyses replaced genetic markers with SNPs to increase resolution for mapping disease loci. Complex diseases are not ideally suited for linkage analyses based on the greater genetic heterogeneity, smaller effect sizes and environmental interactions which confound the strength of potential linkage signals. Furthermore, linkage analyses assume specific modes of inheritance which are not applicable to complex diseases (Pollex and Hegele, 2005).

1.3.2 Population-based techniques

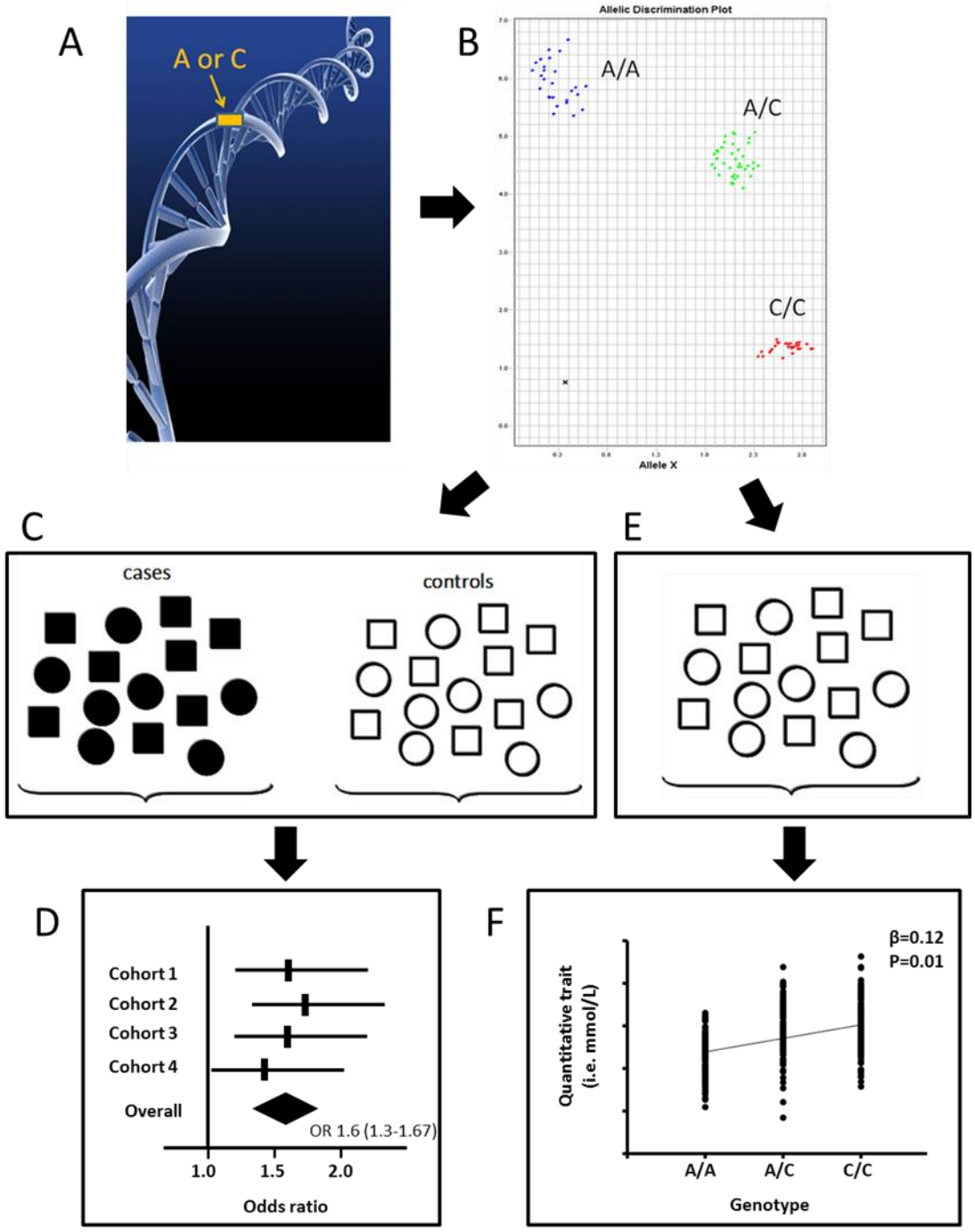
Genetic analyses of the common complex diseases require population-based cohorts where genetic and phenotypic heterogeneity is high. Thus, in the study of complex

disease the traditional approaches used in family-based studies have been replaced by new techniques for identifying disease-related loci. Candidate gene studies represented the first attempt to test for association between genetic variation and disease susceptibility by sequencing genes known *a priori* to be involved in complex disease etiology. Associations identified using this approach suffered from lack of replication in follow-up studies which suggested the need to address the presence of confounding factors.

Genome-wide association studies (GWAS) have become established as an effective and unbiased approach for identifying associations between common genomic variants and complex traits or diseases (Attia et al., 2009; Dube and Hegele, 2012). In this approach, millions of SNPs scattered across the genome are genotyped in a large population using commercially available SNP genotyping arrays. Each individual SNP genotype is then tested for association with discrete case-control status or with a quantitative trait such as plasma cholesterol concentration (**Figure 1.3**). SNP associations are then used as proxies for large haplotype blocks, which implicate a locus with disease susceptibility. As GWAS-identified SNPs most commonly lie in non-coding or intergenic regions, biological relevance is hypothesized based on the genes within the locus with less emphasis on the role of the associated variant. More recently, targeted GWAS have emerged in which custom SNP genotyping arrays are populated with variants that have been previously associated with a disease or trait. The latest targeted GWAS platforms include the Cardio-Metabochip which genotypes common variants associated with various cardio-metabolic traits (Voight et al., 2012). Similarly, the Immunochip

Figure 1.3 Framework for a genome-wide association study. A) The human genome is >99% invariant, but at approximately every 300 nucleotide bases along the DNA string are well-characterized nucleotide sites that vary among people; these sites toggle between 1 of 2 naturally occurring options. For instance, in a pool of individuals, the more common DNA "letter" at a polymorphic site might be "A" for adenosine, and a minority of chromosomes contain the less common form, "C" for cytosine. Such variations are called "single nucleotide polymorphisms" (SNPs). **B)** At each defined SNP site, people have a genotype composed of 2 alleles—1 from each parent—a feature detected with a chemical genotyping method as shown. Each dot on the grid is the SNP genotype for 1 person. Three clusters of people are seen: 2 varieties of homozygotes, namely A/A and C/C, and heterozygotes, namely A/C. It is reasonable to ask whether the members of 1 genotype class differ from those of another. In a candidate gene study, the SNP is deliberately chosen to mark a gene hypothesized a priori to play a mechanistic role determining the trait, but in a genome-wide association study, SNPs from across the entire genome are studied agnostically, without premeditation regarding their possible function. **C)** For case-control studies, SNP genotyping as in **B)** is performed in cases (filled symbols [circles are female, squares are male]) and matched controls. **D)** Odds ratios (ORs) are calculated and displayed on a forest plot: the OR provides an estimate of the risk of conferred by an allele of a given SNP. An allele with an OR >1.0 is associated with increased probability of case status in carriers of the allele and is thus identified as the risk allele. The genotypic ORs for several cohorts can be combined to provide an overall OR for a meta-analysis. **E)** Alternatively, for a quantitative trait association study, SNP genotyping as in **B)** is performed in a sample of the general population. **F)** Linear

regression is used to model the relationship between SNP genotypes and the quantitative trait measurement that provides an estimation of effect size, or β-coefficient, and P-value. The β-coefficient indicates the change in the quantitative trait measurement per copy of the associated allele and is given in the same units as the trait. This figure was reproduced with permission (Dube and Hegele, 2012).


A
A or C
B
Allelic Discrimination Plot
A/A
A/C
C/C
Allele X
C
cases
controls
E
D
Cohort 1
Cohort 2
Cohort 3
Cohort 4
Overall
OR 1.6 (1.3-1.67)
1.0
1.5
2.0
Odds ratio
F
β=0.12
P=0.01
Quantitative trait
(i.e. mmol/L)
A/A
A/C
C/C
Genotype

facilitates the genotyping of variants associated with immunologic pathways (Cortes and Brown, 2011). Since the first published GWAS in 2008, nearly 1500 GWAS have been published on a range of complex diseases, identifying both expected and novel loci in complex disease pathways.

1.3.2.1 Statistics in GWAS

GWAS implementation requires an understanding of several statistical concepts. Firstly, the concept of statistical power is crucial in GWAS design. Statistical power establishes the probability of rejecting the null hypothesis, or no genotype-phenotype association, when no true association exists. Power is calculated based on a given threshold for significance, sample size, the anticipated SNP effect size on disease risk as well as the frequency of the risk-associated variant. Power calculations can help determine the sample size required to detect an association or, conversely, power can be used to assess the effect sizes that can be identified as true association for a given sample size.

Secondly, the tests for association used in GWAS depend upon the phenotype being studied. Discrete phenotypes such as disease status utilize multivariate logistic regression whereas quantitative phenotypes are studied using multivariate linear regression. In either approach, an additive model is calculated to fit the correlation between genotype and phenotype and a P-value is generated which reflects the accuracy of this model. For discrete phenotypes, an odds ratio (OR) is also calculated which provides the frequency ratio for the disease-associated allele in cases versus controls. For

quantitative traits, the calculated β coefficient represents the effect size per allele copy and is given in the units of the trait.

Another important consideration involves the determination of the threshold of significance. Significance, as indicated by the P-value, derived from each test for association provides a measure of the strength of association. Because millions of SNPs are being simultaneously tested for association, the standard false-positive rate of 5% is considered inadequate as a GWAS of one million SNPs will expectedly yield $5x10^{4}$ false positive associations. Thus a Bonferroni-corrected threshold of significance is applied where the standard P-value of 0.05 is divided by the number of SNPs being tested for association. In the case of a GWAS utilizing one million SNPs, the Bonferroni-adjusted P-value would be $5x10^{-8}$; any P-values below this threshold are thus considered statistically significant.

Quality control measures used in GWAS also require statistical context. Particularly in GWAS of discrete phenotypes, it is important to limit the presence of any population substructure or stratification that may create spurious associations. Unequal proportions of individuals from different populations, such as multi-ethnic cohorts, may create differences in allele frequency independent of disease status. In order to address potential population stratification, the case population should not differ significantly from the control population in terms of demographic composition which includes potentially confounding characteristics such as age or sex. It is also preferable to investigate participants of a single ethnicity in order to limit ethnicity-related genetic variation.

Statistically, principal component analysis and genomic control are two widely implemented approaches to address population stratification (Price et al., 2006).

1.3.3 Resequencing studies

In the wake of GWAS, there has been increasing interest in resequencing GWAS-identified loci in order to identify low-frequency variants with potentially larger effect sizes on disease susceptibility. As this approach is focused on rare variants, it will be difficult to test for association in the manner applied with GWAS. Thus studies compare the accumulation of rare variants at targeted loci in case and control populations. Imputation of rare variant genotypes using publicly available GWAS data sets has been used to expand the effective sample size and facilitate association testing between rare variants and disease status (Johansen et al., 2010).

1.3.4 Association studies across ethnicities

Replication of disease-variant associations remains the gold standard for validating GWAS findings. As GWAS have been heavily weighted towards populations of northern European ancestry, replication of the top GWAS findings in multi-ethnic cohorts has become an important stage in GWAS validation (Cooper et al., 2008). The inclusion of well-defined multi-ethnic populations in GWAS incorporates a greater range of human genetic diversity defined by differences in both allele frequencies and LD patterns via varying haplotype block structures. Genotype-phenotype associations that are consistently observed in multi-ethnic studies provide increased confidence behind putative disease loci. Ethnicity-specific differences in allele frequencies, while potentially

confounding, may also help to identify novel disease loci that may not have reached genome-wide significance in other studies simply due to low allele frequencies and thus limited statistical power (Pulit et al., 2010).

1.3.5 Genetic studies in population isolates

Population isolates represent populations that grew from a small group of founders. Due to a combination of geographic or cultural isolation as well as additional forms of genetic drift such as population bottlenecks, genetic diversity within population isolates is significantly lower than the genetic diversity observed in the general population (Arcos-Burgos and Muenke, 2002). Well-known population isolates such as the Old Order Amish and the Finnish have been studied extensively in genetic mapping studies of heritable phenotypes and have revealed several advantages to genetic mapping studies in population isolates (Peltonen et al., 2000). Firstly, the limited genetic heterogeneity within population isolates provides a statistical advantage to detecting association signals as discussed earlier. Given the limited genetic heterogeneity as a result of the founder effect, inbreeding within population isolates is virtually unavoidable and enhances the prevalence of recessive disorders through loss of heterozygosity. Secondly, population isolates are exposed to similar environmental and cultural factors; a crucial characteristic which limits the effect of potentially confounding environmental factors. Thirdly, well-documented multi-generational pedigrees have been documented in many population isolates which facilitates linkage analysis. Population isolates are also well-suited for the study of complex phenotypes for the same reasons that have made them ideal for the study of Mendelian disorders (Kristiansson et al., 2008). The application of GWAS to

population isolates has presented an alternative and statistically favourable method for identifying novel disease loci.

1.4 Genetic architecture of select human diseases

Common complex diseases now account for the leading causes of mortality in Western society and cardiovascular disease (CVD) ranks among the top three. While CVD can be studied using downstream major events, such as myocardial infarction, it is also important to study the genetics of CVD risk factors in order to piece together the various mechanisms that modulate CVD risk. Thus, studies on clinically important variables, such as plasma lipid profile, which is largely determined by genetic factors (Hegele, 2009), have contributed greatly to our understanding of predisposition to CVD. Another highly prevalent CVD co-morbidity in Western society is type 2 diabetes mellitus (T2D). Twin and family studies have similarly ascribed a significant genetic component in T2D susceptibility, which has made T2D the focus of large-scale multi-ethnic GWAS (Imamura and Maeda, 2011). Dementia-related diseases in the elderly are also anticipated to dramatically rise with the aging population, which has inspired intense investigation into the causes of common conditions such as AD which is believed to be considerably heritable. Together, these diseases represent major public health concerns where genetic analyses can contribute significantly to treatment and prevention strategies.

1.4.1 Familial hypercholesterolemia

Familial hypercholesterolemia (FH, OMIM 143890) is an autosomal dominant disease in which dysregulation of low-density lipoprotein (LDL) homeostasis leads to drastically elevated plasma LDL concentrations, and results in premature atherosclerosis and coronary heart disease (CHD) (Liyanage et al., 2011). FH has played an important role in developing our understanding of CVD risk factors, as the markedly elevated LDL cholesterol (LDL-C) levels implicated plasma cholesterol in CVD susceptibility. Subsequent research has characterized the process of atherosclerosis as a complex system involving lipid accumulation in artery walls and chronic inflammation (Lusis, 2000). Through FH studies, it was established that severely elevated plasma cholesterol was sufficient to enhance atherosclerosis, thus establishing plasma cholesterol and LDL-C as robust CVD risk factors in the general population.

1.4.1.1 Pathophysiology and genetic architecture

Clinical diagnosis of FH varies but typically includes childhood presentation of xanthomas and LDL cholesterol (LDL-C) >95th percentile (Raal and Santos, 2012). As described by Brown and Goldstein in their Nobel Prize-winning research (Goldstein and Brown, 2009), the elevated plasma LDL-C observed in FH is due to variation in the gene encoding the low-density lipoprotein receptor (LDLR); a cell surface receptor that binds and internalizes circulating LDL particles. Rare loss-of-function *LDLR* mutations impair plasma LDL homeostasis leading to hypercholesterolemia and the accelerated formation of atherosclerotic plaques. While mutations in *LDLR* account for the majority of FH cases, mutations in the *APOB* and *PCSK9* genes also produce similar

hypercholesterolemia phenotypes (Raal and Santos, 2012). Interestingly, heterozygous FH (HeFH) is quite common for a monogenic disease. HeFH is generally observed at a frequency of 1:500 but has been reported at much higher frequencies in specific founder populations such as Dutch South Africans where HeFH frequency is reportedly 1:70; the frequency of homozygous FH (HoFH) in the general population is considerably lower at $1:10^6$ (Liyanage et al., 2011).

1.4.1.2 Plasma cholesterol as a complex trait

The genetic basis for most Mendelian dyslipidemias such as FH has been solved while our understanding of the genetic variation underlying common lipid trait variance in the general population remains incomplete. Through the application of GWAS to lipoprotein concentration in population-based samples, genes associated with Mendelian dyslipidemias have also been associated with the variance in lipid concentration observed in the general population. For example, common variants in *LDLR, APOB* and *PCSK9* have all been highly associated with plasma LDL concentration in the largest GWAS meta-analysis on plasma lipid traits (Teslovich et al., 2010). Perhaps more interestingly, novel and unexpected LDL-C-associated loci have emerged such as the *SORT1* locus which has subsequently been validated as a mediator of plasma LDL concentration by binding and internalizing circulating apoB-containing lipoproteins (Dube et al., 2011). A major caveat, however, relates to the fact that only a modest portion of variance in LDL concentration has been attributed to common variation (Willer and Mohlke, 2012). While a portion of the body's cholesterol is derived from the environment, the overwhelming portion of the cholesterol pool is synthesized endogenously and is thus anticipated to be

largely genetically determined (Hegele, 2009). One estimate ascribed 50% heritability to plasma LDL concentration which can be contrasted against the ~25-30% variance explained by genetic variation at 95 loci (Perusse et al., 1997; Teslovich et al., 2010). Resequencing of GWAS-identified genes for the detection of additional common and rare variants has been shown to increase the portion of explained heritability in LDL concentration and supports the execution of large-scale whole genome sequencing efforts in the future (Sanna et al., 2011).

1.4.2 Late-onset cognitive decline and dementia

More than 300 psychiatric disorders have been described where an established mechanism of pathogenesis is often absent (Sullivan et al., 2012). With the growing elderly population, there has been concern regarding the anticipated rise in geriatric psychiatric disorders, our ability to treat these disorders as well as the greater burden that will be placed on already exhausted healthcare expenditure. By studying the genetics of heritable late-onset psychiatric disorders, it may be possible to dissect the biological pathways that modulate a phenotype as complex as cognitive health and to develop early diagnostic and intervention strategies.

1.4.2.1 Spectrum of disease severity

Dementia represents a common end-point for many psychiatric disorders and affects ~5% of the elderly (Eaton et al., 2008). Accordingly, the clinical definition broadly includes impairment in memory as well as at least one other cognitive domain which includes language, calculations, orientation and judgment. Importantly, the cognitive loss must be

of sufficient severity as to significantly disable social or occupational autonomy. Cognitive testing scores are also used to measure and track the state of the patient's cognitive health (Kawas, 2003). Although cognitive decline is associated with normal aging, the observed cognitive deficits are not severe enough to interfere with the patient's autonomy. As cognitive impairment in the elderly is considered to be a degenerative process, it is widely believed that a period of intermediate yet measurable cognitive impairment precedes dementia and has been described as "cognitive impairment, no dementia" (CIND) or mild cognitive impairment (MCI) (Graham et al., 1997). The major distinction between CIND and MCI involves the more inclusive definition of CIND which is based on the exclusion of dementia and clinical evidence of any form of cognitive impairment (Graham et al., 1997); MCI is specific to pre-dementia where AD is suspected (Voisin et al., 2003). Both CIND and MCI have frequencies of ~20% among the elderly with CIND patients 5 times more likely to develop dementia while 10% - 15% of MCI patients progress to AD (Tarawneh and Holtzman, 2012; Tuokko et al., 2003). Logistically, pre-dementia studies have been fraught by phenotypic heterogeneity as patients can progress to dementia, remain stable or improve cognitively.

1.4.2.2 Alzheimer disease

AD is by far the most prevalent cognitive disease affecting elderly people >65 years old. Among the elderly, AD underlies >70% of dementia cases and represents the sixth highest cause of death across all ages in the United States (Tarawneh and Holtzman, 2012). Twin studies have also ascribed considerable heritability in AD with estimates ranging from 58% - 79% (Gatz et al., 2006). Genetic studies in AD hit a major

breakthrough when autosomal dominant mutations in amyloid beta precursor protein (*APP*), presenilin 1 (*PSEN1*) and 2 (*PSEN2*) were shown to cause the rare early-onset form of AD (Tanzi, 2012). GWAS on the more prevalent late-onset form of AD identified apolipoprotein E (*APOE*) as a candidate gene. Additional susceptibility loci implicated immune and inflammatory pathways as well as lipid trafficking pathways. Cumulatively, common variation is estimated to explain ~33% of the genetic risk. Based on GWAS findings, the potential roles of immunity, inflammation and lipid metabolism in AD pathogenesis provide new perspectives to approach AD risk which may help inform future genetic strategies for studying AD.

1.4.2.3 Vascular dementia

Vascular dementia (VaD) refers to dementia resulting from cerebrovascular dysfunction and accounts for ~16% of dementia cases in the elderly (Ott et al., 1995). Cerebrovascular dysfunction can manifest as subclinical brain injury, silent brain infarction or stroke; all of which contribute toward a damaging environment of chronic ischemia in the brain. Twin studies have shown little support for a significant heritable component in VaD and, perhaps not surprisingly, GWAS on VaD have failed to return genome-wide significant results. Although genetic approaches to studying VaD have not been encouraging, they have also been hampered by the phenotypic heterogeneity underlying VaD. Additionally, alternative approaches may be useful in studying VaD susceptibility which includes investigating the genetics of known VaD risk factors. It is also well-known that the cardio-metabolic dysfunction that enhances atherosclerosis and CVD in the periphery correlates with cognitive health in that CVD risk factors such as

lipid profile, obesity and T2D have each been linked with enhanced risk of cognitive decline (Gorelick et al., 2011). Thus an approach focused on genetic variation at cardio-metabolic loci in VaD may shed new light on predisposition to cognitive decline.

1.4.3 Type 2 diabetes

Diabetes broadly refers to a group of metabolic diseases defined by abnormal plasma glucose homeostasis. T2D is the most prevalent form of diabetes and is clinically characterized by hyperglycemia, insulin resistance and impaired insulin secretion (Patel and Macerollo, 2010). According to a World Health Organization study, T2D had a worldwide prevalence of 2.8% across all age groups in 2000, which accounted for 171 million people (Wild et al., 2004). T2D has devastating long-term effects on the body such as macrovascular and microvascular complications as well as effects on lipid homeostasis. Accordingly, T2D has been strongly implicated as an independent risk factor for CVD (Wilson, 1998; Wilson et al., 1998). As a complex disease, T2D risk is in part mediated by lifestyle and environmental factors which can be managed. The genetic aspect of T2D susceptibility, however, has helped reveal some of the biological pathways underlying T2D pathogenesis which may be targeted for therapeutic intervention. Considering that ~65% of diabetics die from CVD-related causes, improved understanding and treatment of T2D may have a significant impact on improving patient quality of life (Grundy et al., 1999).

1.4.3.1 Pathophysiology

T2D pathogenesis is driven largely by 1) dysfunctional pancreatic β cells which produce and secrete insulin; 2) excess hepatic glucose production; and 3) insulin resistance where insulin-mediated glucose clearance is disrupted (Leahy, 2005). Insulin represents the key hormone responsible for maintaining glycemic control where insulin secretion supports normoglycemia (Stumvoll et al., 2005). Prior to T2D, patients develop progressive insulin resistance where the glucose-lowering effects of insulin are gradually impaired and normal glycemic homeostasis is disrupted leading to hyperglycemia. β cell function ramps up in order to compensate for the insulin insensitivity which leads to observed hyperinsulinemia. With chronic and worsening hyperglycemia, the β cell cannot maintain the high rate of insulin secretion. Thus β cell function and mass both deteriorate as the disease progresses which further contributes to the dysregulation of glycemic homeostasis (Leahy, 2005). Together, the chronic dysregulation of glycemic homeostasis is manifested as major complications such as diabetic retinopathy, nephropathy and neuropathy.

1.4.3.2 Common genetic risk factors

As with many common complex diseases, family and twin studies suggested a strong heritable component in T2D susceptibility (Ahlqvist et al., 2011). Furthermore, T2D rates were observed to vary significantly between ethnic groups living within the same environment where African Americans, Hispanic Americans and aboriginal North Americans are at greater risk of T2D than American Caucasians (Carter et al., 1996; Harris et al., 1998). Although linkage mapping studies first proposed chromosomal

regions potentially housing T2D susceptibility genes, it was not until the high-resolution GWAS approach was applied that T2D loci were discretely identified. The largest meta-analyses of T2D GWAS data have revealed as many as 58 T2D susceptibility loci which strongly implicate the influence of β cell function at the centre of T2D susceptibility (Voight et al., 2010; Zeggini et al., 2008). Despite these successes in expanding our understanding of T2D-related pathways, common variation accounts for only ~10% of T2D heritability. Clearly, further investigation is required in order to better understand the nature of T2D risk and to account for the missing heritability.

1.4.3.3 Prevalence within First Nations communities

North American aboriginal populations are at increased T2D risk compared to the non-native population. For example, T2D affects ~9% of the general Canadian population versus 26% of Oji-Cree in Ontario, Canada (Lipscombe and Hux, 2007; Yu and Zinman, 2007). This increased T2D risk among the First Nations also extends to aboriginal children as the incidence of childhood T2D in Manitoba first nations was nearly 8 times higher compared to children across Canada in 2004-2006 (Amed et al., 2010). Interestingly, Inuit populations in Canada and Alaska have shown low T2D prevalence, however, high impaired glucose tolerance (IGT) prevalence has suggested that T2D may eventually become problematic among Inuit populations (Pedersen, 2012). The enhanced T2D risk among certain aboriginal Canadians identified epidemiologically is likely to be explained by environmental or genetic components unique to these populations. Genetic studies in aboriginal groups have been limited, however, a candidate gene study of *HNF1A* in an Oji-Cree population identified the private p.G319S variant which was

present in 40% of the population and increased T2D risk by 4-fold (Hegele et al., 1999). The discovery of a common large-effect T2D variant was crucial in understanding the propagation of T2D among the Oji-Cree and also underscored the importance of investigating population-specific risk factors in understanding disease pathogenesis. As genetic studies of aboriginal populations are relatively few, the investigation of putative T2D-associated variation in isolated populations such as the Oji-Cree may help in establishing T2D risk factors across ethnicities.

1.5 Summary

The concept of common and complex human disease has changed due to technological and practical innovations in science and medicine. Although infectious diseases no longer devastate Western society, the overall increase in life expectancy has led to the emergence of complex diseases as the current leading causes of mortality. Common complex disease etiology has been difficult to characterize since an intricate synergy between environmental, epigenetic and genetic risk factors is suspected. Thus considerable heterogeneity exists between any two patients diagnosed with CVD due to the fact that, while their disease endpoints may be similar, the respective collection of risk factors may be disparate. But while environmental factors may be modified, our genetic make-up for the most part remains stable and so the hunt for robust genetic determinants of complex disease susceptibility has been a major focus emerging from the genomics era.

The overall hypothesis of experiments performed in this thesis is that common genetic variation contributes to the variance observed in common complex diseases and traits and can be used to study complex disease susceptibility in multi-ethnic cohorts. We sought to test this hypothesis through genetic analyses of three distinct complex disease-related phenotypes with the following objectives: 1) conduct a resequencing study of *LDLR* in genetically isolated Inuit populations to identify novel variants associated with LDL cholesterol and CVD risk; 2) conduct a targeted GWAS of common variation previously linked with cardio-metabolic traits and AD in CIND patients in order to evaluate the role of these pathways in CIND susceptibility; and 3) conduct a candidate genotyping study of GWAS-identified T2D variants in Canadian aboriginal populations in order to evaluate the applicability of established T2D loci in multi-ethnic populations. Together, these studies exemplify different approaches to study the genetic basis of complex disease and provide analytical workflows which may help in the design of future studies on complex disease. As the genomics era is rapidly evolving with the introduction of feasible whole genome sequencing, the concepts and analyses discussed in this thesis are likely to remain fundamental to the future study of complex disease genetics.

1.6 References

Ahlqvist, E., Ahluwalia, T.S., and Groop, L. (2011). Genetics of type 2 diabetes. *Clin Chem 57*, 241-254.

Amed, S., Dean, H.J., Panagiotopoulos, C., Sellers, E.A., Hadjiyannakis, S., Laubscher, T.A., Dannenbaum, D., Shah, B.R., Booth, G.L., and Hamilton, J.K. (2010). Type 2 diabetes, medication-induced diabetes, and monogenic diabetes in Canadian children: a prospective national surveillance study. *Diabetes Care 33*, 786-791.

Apostolou, P., and Fostira, F. (2013). Hereditary breast cancer: the era of new susceptibility genes. *Biomed Res Int 2013*, 747318.

Arcos-Burgos, M., and Muenke, M. (2002). Genetics of population isolates. *Clin Genet 61*, 233-247.

Attia, J., Ioannidis, J.P., Thakkinstian, A., McEvoy, M., Scott, R.J., Minelli, C., Thompson, J., Infante-Rivard, C., and Guyatt, G. (2009). How to use an article about genetic association: A: Background concepts. *JAMA 301*, 74-81.

Carter, J.S., Pugh, J.A., and Monterrosa, A. (1996). Non-insulin-dependent diabetes mellitus in minorities in the United States. *Ann Intern Med 125*, 221-232.

Centers for Disease Control and Prevention, Leading Causes of Death, 1900-1998, [March 1, 2013 accessed].

Cooper, R.S., Tayo, B., and Zhu, X. (2008). Genome-wide association studies: implications for multiethnic samples. *Hum Mol Genet 17*, R151-155.

Cortes, A., and Brown, M.A. (2011). Promise and pitfalls of the Immunochip. *Arthritis Res Ther 13*, 101.

Dawn Teare, M., and Barrett, J.H. (2005). Genetic linkage studies. *Lancet 366*, 1036-1044.

Donoghue, H.D. (2009). Human tuberculosis--an ancient disease, as elucidated by ancient microbial biomolecules. *Microbes Infect 11*, 1156-1162.

Dube, J.B., and Hegele, R.A. (2012). Genetics 100 for cardiologists: basics of genome-wide association studies. *Can J Cardiol 29*, 10-17.

Dube, J.B., Johansen, C.T., and Hegele, R.A. (2011). Sortilin: an unusual suspect in cholesterol metabolism: from GWAS identification to in vivo biochemical analyses, sortilin has been identified as a novel mediator of human lipoprotein metabolism. *Bioessays 33*, 430-437.

Eaton, W.W., Martins, S.S., Nestadt, G., Bienvenu, O.J., Clarke, D., and Alexandre, P. (2008). The burden of mental disorders. *Epidemiol Rev 30*, 1-14.

Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., *et al.* (2002). The structure of haplotype blocks in the human genome. *Science 296*, 2225-2229.

Gatz, M., Reynolds, C.A., Fratiglioni, L., Johansson, B., Mortimer, J.A., Berg, S., Fiske, A., and Pedersen, N.L. (2006). Role of genes and environments for explaining Alzheimer disease. *Arch Gen Psychiatry 63*, 168-174.

Goldstein, J.L., and Brown, M.S. (2009). The LDL receptor. *Arterioscler Thromb Vasc Biol 29*, 431-438.

Gorelick, P.B., Scuteri, A., Black, S.E., Decarli, C., Greenberg, S.M., Iadecola, C., Launer, L.J., Laurent, S., Lopez, O.L., Nyenhuis, D., *et al.* (2011). Vascular contributions to cognitive impairment and dementia: a statement for healthcare professionals from the american heart association/american stroke association. *Stroke 42*, 2672-2713.

Graham, J.E., Rockwood, K., Beattie, B.L., Eastwood, R., Gauthier, S., Tuokko, H., and McDowell, I. (1997). Prevalence and severity of cognitive impairment with and without dementia in an elderly population. *Lancet 349*, 1793-1796.

Grundy, S.M., Benjamin, I.J., Burke, G.L., Chait, A., Eckel, R.H., Howard, B.V., Mitch, W., Smith, S.C., Jr., and Sowers, J.R. (1999). Diabetes and cardiovascular disease: a statement for healthcare professionals from the American Heart Association. *Circulation 100*, 1134-1146.

Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A., and McKusick, V.A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res 33*, D514-517.

Harris, M.I., Flegal, K.M., Cowie, C.C., Eberhardt, M.S., Goldstein, D.E., Little, R.R., Wiedmeyer, H.M., and Byrd-Holt, D.D. (1998). Prevalence of diabetes, impaired fasting glucose, and impaired glucose tolerance in U.S. adults. The Third National Health and Nutrition Examination Survey, 1988-1994. *Diabetes Care 21*, 518-524.

Hegele, R.A. (2009). Plasma lipoproteins: genetic influences and clinical implications. *Nat Rev Genet 10*, 109-121.

Hegele, R.A., Cao, H., Harris, S.B., Hanley, A.J., and Zinman, B. (1999). The hepatic nuclear factor-1alpha G319S variant is associated with early-onset type 2 diabetes in Canadian Oji-Cree. *J Clin Endocrinol Metab 84*, 1077-1082.

Heron, M. (2012). Deaths: leading causes for 2008. *Natl Vital Stat Rep 60*, 1-94.

Imamura, M., and Maeda, S. (2011). Genetics of type 2 diabetes: the GWAS era and future perspectives [Review]. *Endocr J 58*, 723-739.

International HapMap Consortium, T. (2003). The International HapMap Project. *Nature 426*, 789-796.

Johansen, C.T., Wang, J., Lanktree, M.B., Cao, H., McIntyre, A.D., Ban, M.R., Martins, R.A., Kennedy, B.A., Hassell, R.G., Visser, M.E., *et al.* (2010). Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. *Nat Genet 42*, 684-687.

Kawas, C.H. (2003). Clinical practice. Early Alzheimer's disease. *N Engl J Med 349*, 1056-1063.

Kristiansson, K., Naukkarinen, J., and Peltonen, L. (2008). Isolated populations and complex disease gene identification. *Genome Biol 9*, 109.

Leahy, J.L. (2005). Pathogenesis of type 2 diabetes mellitus. *Arch Med Res 36*, 197-209.

Ligon, B.L. (2006). Plague: a review of its history and potential as a biological weapon. *Semin Pediatr Infect Dis 17*, 161-170.

Lipscombe, L.L., and Hux, J.E. (2007). Trends in diabetes prevalence, incidence, and mortality in Ontario, Canada 1995-2005: a population-based study. *Lancet 369*, 750-756.

Liyanage, K.E., Burnett, J.R., Hooper, A.J., and van Bockxmeer, F.M. (2011). Familial hypercholesterolemia: epidemiology, Neolithic origins and modern geographic distribution. *Crit Rev Clin Lab Sci 48*, 1-18.

Lozano, R., Naghavi, M., Foreman, K., Lim, S., Shibuya, K., Aboyans, V., Abraham, J., Adair, T., Aggarwal, R., Ahn, S.Y., *et al.* (2013). Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet 380*, 2095-2128.

Lusis, A.J. (2000). Atherosclerosis. *Nature 407*, 233-241.

Mangino, M., and Spector, T. (2012). Understanding coronary artery disease using twin studies. *Heart*.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., *et al.* (2009). Finding the missing heritability of complex diseases. *Nature 461*, 747-753.

Mayo, O. (2008). A century of Hardy-Weinberg equilibrium. *Twin Res Hum Genet 11*, 249-256.

Ott, A., Breteler, M.M., van Harskamp, F., Claus, J.J., van der Cammen, T.J., Grobbee, D.E., and Hofman, A. (1995). Prevalence of Alzheimer's disease and vascular dementia: association with education. The Rotterdam study. *BMJ 310*, 970-973.

Patel, P., and Macerollo, A. (2010). Diabetes mellitus: diagnosis and screening. *Am Fam Physician 81*, 863-870.

Pearson, T.A., and Manolio, T.A. (2008). How to interpret a genome-wide association study. *JAMA 299*, 1335-1344.

Pedersen, M.L. (2012). Diabetes mellitus in Greenland. *Dan Med J 59*, B4386.

Peltonen, L., Palotie, A., and Lange, K. (2000). Use of population isolates for mapping complex traits. *Nat Rev Genet 1*, 182-190.

Perusse, L., Rice, T., Despres, J.P., Bergeron, J., Province, M.A., Gagnon, J., Leon, A.S., Rao, D.C., Skinner, J.S., Wilmore, J.H., *et al.* (1997). Familial resemblance of plasma lipids, lipoproteins and postheparin lipoprotein and hepatic lipases in the HERITAGE Family Study. *Arterioscler Thromb Vasc Biol 17*, 3263-3269.

Pollex, R.L., and Hegele, R.A. (2005). Complex trait locus linkage mapping in atherosclerosis: time to take a step back before moving forward? *Arterioscler Thromb Vasc Biol 25*, 1541-1544.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet 38*, 904-909.

Pritchard, J.K. (2001). Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet 69*, 124-137.

Pulit, S.L., Voight, B.F., and de Bakker, P.I. (2010). Multiethnic genetic association studies improve power for locus discovery. *PLoS One 5*, e12600.

Raal, F.J., and Santos, R.D. (2012). Homozygous familial hypercholesterolemia: current perspectives on diagnosis and treatment. *Atherosclerosis 223*, 262-268.

Reich, D.E., and Lander, E.S. (2001). On the allelic spectrum of human disease. *Trends Genet 17*, 502-510.

Sanna, S., Li, B., Mulas, A., Sidore, C., Kang, H.M., Jackson, A.U., Piras, M.G., Usala, G., Maninchedda, G., Sassu, A., *et al.* (2011). Fine mapping of five loci associated with low-density lipoprotein cholesterol detects variants that double the explained heritability. *PLoS Genet 7*, e1002198.

Stumvoll, M., Goldstein, B.J., and van Haeften, T.W. (2005). Type 2 diabetes: principles of pathogenesis and therapy. *Lancet 365*, 1333-1346.

Sullivan, P.F., Daly, M.J., and O'Donovan, M. (2012). Genetic architectures of psychiatric disorders: the emerging picture and its implications. *Nat Rev Genet 13*, 537-551.

Tanzi, R.E. (2012). The genetics of Alzheimer disease. *Cold Spring Harb Perspect Med 2*.

Tarawneh, R., and Holtzman, D.M. (2012). The clinical problem of symptomatic Alzheimer disease and mild cognitive impairment. *Cold Spring Harb Perspect Med 2*, a006148.

Teslovich, T.M., Musunuru, K., Smith, A.V., Edmondson, A.C., Stylianou, I.M., Koseki, M., Pirruccello, J.P., Ripatti, S., Chasman, D.I., Willer, C.J., *et al.* (2010). Biological, clinical and population relevance of 95 loci for blood lipids. *Nature 466*, 707-713.

Tishkoff, S.A., and Kidd, K.K. (2004). Implications of biogeography of human populations for 'race' and medicine. *Nat Genet 36*, S21-27.

Tuokko, H., Frerichs, R., Graham, J., Rockwood, K., Kristjansson, B., Fisk, J., Bergman, H., Kozma, A., and McDowell, I. (2003). Five-year follow-up of cognitive impairment with no dementia. *Arch Neurol 60*, 577-582.

Varki, A., and Altheide, T.K. (2005). Comparing the human and chimpanzee genomes: searching for needles in a haystack. *Genome Res 15*, 1746-1758.

Voight, B.F., Kang, H.M., Ding, J., Palmer, C.D., Sidore, C., Chines, P.S., Burtt, N.P., Fuchsberger, C., Li, Y., Erdmann, J., *et al.* (2012). The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS Genet 8*, e1002793.

Voight, B.F., Scott, L.J., Steinthorsdottir, V., Morris, A.P., Dina, C., Welch, R.P., Zeggini, E., Huth, C., Aulchenko, Y.S., Thorleifsson, G., *et al.* (2010). Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat Genet 42*, 579-589.

Voisin, T., Touchon, J., and Vellas, B. (2003). Mild cognitive impairment: a nosological entity? *Curr Opin Neurol 16 Suppl 2*, S43-45.

Wild, S., Roglic, G., Green, A., Sicree, R., and King, H. (2004). Global prevalence of diabetes: estimates for the year 2000 and projections for 2030. *Diabetes Care 27*, 1047-1053.

Willer, C.J., and Mohlke, K.L. (2012). Finding genes and variants for lipid levels after genome-wide association analysis. *Curr Opin Lipidol 23*, 98-103.

Wilson, P.W. (1998). Diabetes mellitus and coronary heart disease. *Am J Kidney Dis 32*, S89-100.

Wilson, P.W., D'Agostino, R.B., Levy, D., Belanger, A.M., Silbershatz, H., and Kannel, W.B. (1998). Prediction of coronary heart disease using risk factor categories. *Circulation 97*, 1837-1847.

Yu, C.H., and Zinman, B. (2007). Type 2 diabetes and impaired glucose tolerance in aboriginal populations: a global perspective. *Diabetes Res Clin Pract 78*, 159-170.

Zeggini, E., Scott, L.J., Saxena, R., Voight, B.F., Marchini, J.L., Hu, T., de Bakker, P.I., Abecasis, G.R., Almgren, P., Andersen, G., *et al.* (2008). Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat Genet 40*, 638-645.

CHAPTER 2

THE PRIVATE, COMMON LDLR p.G116S VARIANT HAS A LARGE EFFECT ON PLASMA LDL CHOLESTEROL IN CIRCUMPOLAR POPULATIONS

2.1 INTRODUCTION

Cardiovascular disease (CVD) represents a complex condition marked by progressive atherosclerosis and inflammation leading to arterial occlusion and myocardial infarction (Roy et al., 2009). Plasma low-density lipoprotein (LDL) concentration and, more specifically, LDL cholesterol (LDL-C) represent major risk factors in determining CVD risk. The lowering of LDL-C remains a primary goal in clinical CVD management (Genest et al., 2009; Gotto and Moon, 2012). Over the past few decades, overall mortality rates due to CVD have been on the decline particularly in North America, which is believed to be due in large part to improved management of risk factors and improved therapeutic interventions (Carroll et al., 2012; Gregg et al., 2005). Despite the overall lower CVD risk, a countercurrent trend has emerged within global aboriginal and indigenous communities where the increasing Westernization of diet and lifestyle has correlated with increased prevalence of type 2 diabetes, obesity and ultimately CVD risk (Stoner et al., 2012; Yu and Zinman, 2007).

Among the northerly aboriginal populations, Inuit descendants have presented a unique case with respect to CVD risk as, historically, it was believed that Inuit descendants were at lower CVD risk than non-native populations (Bjerregaard and

Dyerberg, 1988; Middaugh, 1990; Young et al., 1993). It was a commonly held belief that CVD was virtually non-existent within Inuit communities based on cardio-protective effects of the traditional lifestyle and marine diet (Dewailly et al., 2001). Genetic factors were also considered to account for the apparent cardiovascular protection observed among the Inuit, however, a study on candidate CVD-associated variants among Inuit of the Keewatin region of modern Nunavut showed that the Inuit carried a higher frequency of certain CVD-associated variants compared to the non-native population (Hegele et al., 1997). A closer analysis of CVD studies in Inuit populations suggested that the data used to establish the concept of cardiovascular protection among the Inuit relative to non-native populations was likely unfounded and that unreliable mortality statistics may have helped perpetuate this concept when, in fact, rates of ischemic heart disease were similar to non-native populations (Bjerregaard et al., 2003b). Preventive action against CVD within Inuit communities has thus become increasingly paramount particularly as progressive Westernization has gradually ushered in CVD risk factors such as smoking, higher caloric intake and sedentary lifestyle that, while not immediately detrimental, may potentially affect CVD prevalence for future generations (Bjerregaard et al., 1997; Bjerregaard et al., 2003b; Chateau-Degat et al., 2010; Ebbesson et al., 2005; Howard et al., 2010; Jernigan et al., 2010; Kellett et al., 2012).

As re-evaluation of Inuit health statistics show that the Inuit are not uniquely protected from CVD risk (Schumacher et al., 2003), it is important to consider the unique risk factors to which Inuit populations may be exposed. Recent Inuit health studies have shown that a large percentage of adults live with high LDL-C (Jorgensen et al., 2008;

Redwood et al., 2010). The prevalence of LDL-related diseases among the Inuit, particularly familial hypercholesterolemia (FH, Online Mendelian Inheritance in Man [OMIM] 143890) remains unreported and unexplored despite heterozygous FH (HeFH) having one of the highest frequencies for a monogenic disease in North American and European populations at a rate of 1:500 (Haase and Goldberg, 2012).

While LDL-C can be managed through diet and lifestyle, almost 80% of the body's cholesterol is derived endogenously which has placed greater emphasis on understanding the biological mechanisms that modulate cholesterol homeostasis (Hegele, 2009). Genetic studies on FH identified deleterious mutations in the low-density lipoprotein receptor (LDLR) gene as the cause of the observed hypercholesterolemia and thus implicated LDLR as a major regulator of plasma LDL homeostasis (Goldstein and Brown, 2009). To date, >1,700 hypercholesterolemia-associated variants in *LDLR* have been reported suggesting that the *LDLR* locus is a hotspot for genetic variation. The role of variation in *LDLR* in modulating LDL-C within Inuit populations has not been investigated but given the prevalence of elevated LDL-C among select Inuit cohorts (Bjerregaard et al., 2004; Ebbesson et al., 1996; Redwood et al., 2010), genetic variation may be contributing to the variance in LDL-C observed in Inuit populations and may help identify individuals at elevated CVD risk.

We therefore sought to investigate genetic variation at the *LDLR* locus within Inuit descendants and test for association with lipid traits. Through Sanger sequencing of *LDLR* and targeted genotyping, we report the discovery of two private, common *LDLR*

variants in five Inuit populations from North America and Greenland. The first variant encodes a glycine-to-serine substitution at the 116th amino acid (p.G116S), which was previously reported in a Danish population (Damgaard et al., 2005), and the second variant encodes a novel arginine-to-tryptophan substitution at the 730th amino acid (p.R730W). Subsequent tests for association between these variants and lipid traits strongly associate p.G116S with a large effect on LDL-C while p.R730W had a non-significant effect on LDL-C.

2.2 MATERIALS AND METHODS

2.2.1 Study populations

Participants of Inuit descent and aged >18 years were obtained from various arctic regions within North America and Greenland (**Figure 2.1, Table 2.1**). Within Canada, population-based samples were obtained from the "*Qanuippitaa*" health survey which included 14 coastal communities in Nunavik, Quebec (n=450); the Keewatin Health Assessment Study which surveyed residents of the Keewatin region in Nunavut (n=214) (Moffatt et al., 1993); and the Inuvik region in the North West Territories (n=281). Lastly, we included in our study a sample of Inuit living in Denmark and West Greenland (n=1191) derived from a regional health survey (Bjerregaard et al., 1997) as well as Yup'ik and Cup'ik Inuit from southwestern Alaska (n=1223) as part of the Center for Alaska Native Health Research iniative (Boyer et al., 2007). Population-based samples of the Oji-Cree (n=137) from Manitoulin Island, Ontario were also genotyped for G116S and R730W.

Figure 2.1 A map of select Inuit settlements across North America and Greenland. **A)** Southwest Alaska is home to >20,000 Alaska Natives living in ~50 rural villages of ~500 inhabitants per village where Yup'ik and Cup'ik comprise the major Inuit subpopulations. **B)** The Inuvik region of the Northwest Territories is home to ~3,200 Inuit descendants who represent ~35% of the population (Statistics Canada, 2006). **C)** The former Keewatin region – now known as Kivalliq – in Nunavut is home to ~7,445 Inuit descendants who represent ~89% of the population (Moffatt et al., 1993). **D)** The Nunavik territory of northern Québec lies north of the 55th parallel where the population of ~11,000 is represented by ~91% Inuit. The entire population is spread across fourteen coastal settlements (Counil et al., 2009). **E)** Greenland's population of ~56,000 is represented by ~90% Inuit where the majority of the population lives on the southwestern coast (Bjerregaard et al., 2003a).

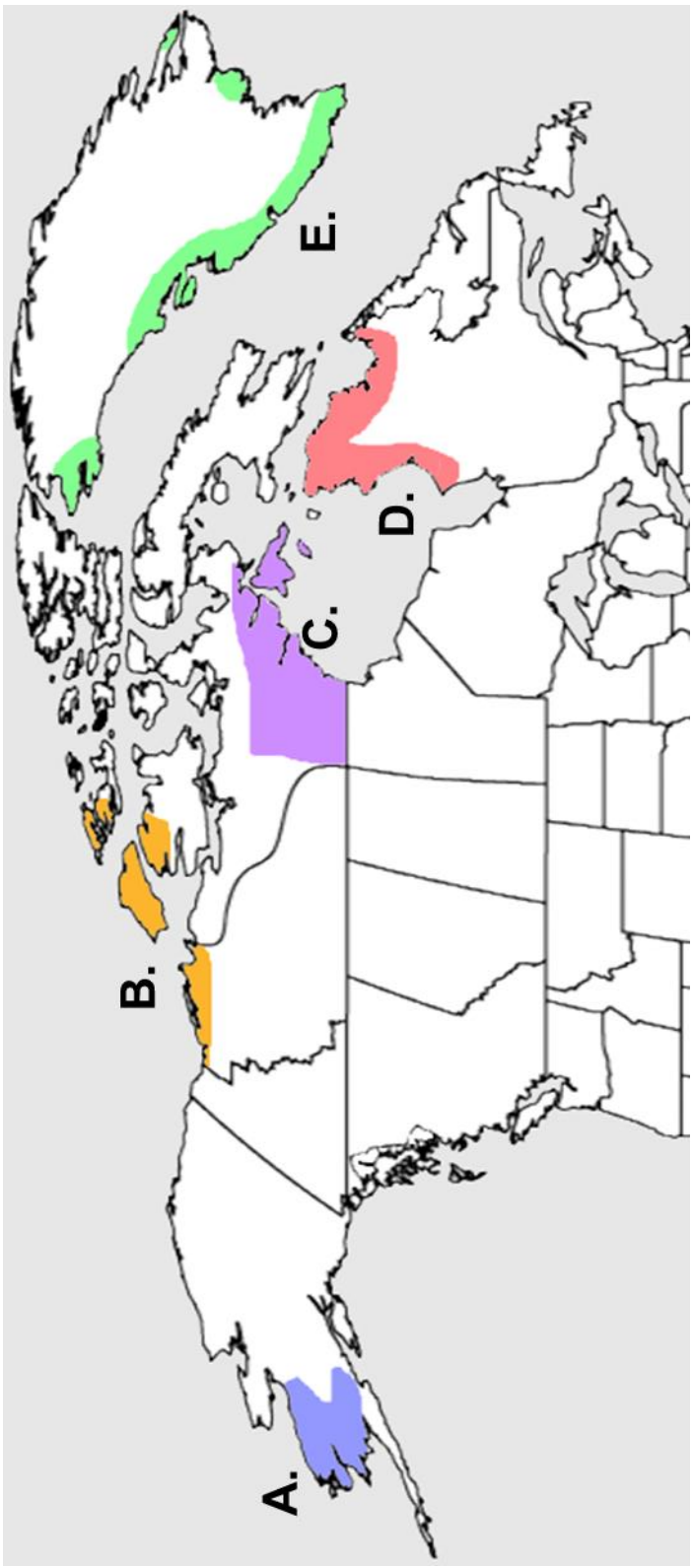

E.
D.
C.
B.
A.

Table 2.1. Demographics and *LDLR* variant frequencies for select circumpolar populations.

	n	Age	Female (%)	BMI (kg/m^2)	TC	LDL-C	HDL-C	Non-HDL-C	apoB	TG	MAF	
											p.G116S	p.R730W
Greenland	1182	44±14	56	26±5	5.91±1.13	3.82±1.04	1.57±0.44	4.33±1.14	0.92±0.23	1.16±0.67	0.13	0.11
Keewatin	210	37±16	54	27±4	5.00±1.03	3.09±0.92	1.45±0.41	3.55±1.01	0.98±0.26	1.03±0.57	0.02	0.17
Inuvik	281	45±16	67	30±7	5.05±0.99	2.91±0.89	1.37±0.42	3.68±1.03	0.91±0.25	1.74±1.27	0.05	0.13
Nunavik	429	37±14	56	27±6	4.99±0.99	2.79±0.86	1.63±0.43	3.33±1.02	0.96±0.24	1.23±0.72	0.09	0.13
Alaska	1222	38±16	53	28±6	5.20±1.15	3.20±0.98	1.64±0.44	3.61±1.08	n.d.	0.94±0.56	0.10	0.16
Combined	3324	40±16	56	27±6	5.40±1.16	3.30±1.05	1.58±0.44	3.83±1.15	0.93±0.24	1.13±0.74	0.10	0.14

All demographics are reported ± standard deviation. Lipid-related traits are all reported in mmol/L except apoB which is in g/L. BMI, body mass index; LDL-C, low-density lipoprotein cholesterol; HDL-C, high-density lipoprotein cholesterol; MAF, minor allele frequency; n.d., no data; TC, total cholesterol concentration; TG, triglyceride concentration.

Generally, participants were asked to fast either for 12 hours or overnight prior to blood sample collection. Plasma lipid concentrations were determined using varying methods across the different Inuit population studies. Carotid intima-media thickness (IMT) measurements were obtained using established ultrasound protocols. Measurements were calculated for 12 1-cm segments which included the near and far walls of the common carotid artery and bifurcation of the common carotid artery. For analysis with *LDLR* variant genotypes, mean IMT included an average of all 12 segments; mean common carotid IMT (C-IMT) included the average of 4 segments exclusively covering the common carotid artery.

2.2.2 Study design

The *LDLR* promoter region and exons were Sanger sequenced within a subset of Greenland Inuit participants (n=10) with LDL-C concentrations >95th percentile (~6.00 mmol/L). The p.G116S and p.R730W variants were identified in exons 4 and 15 of *LDLR* respectively (**Figure 2.2**). Study participants from four independent Inuit populations were subsequently genotyped for the two variants (**Table 2.2**). Genotypes were first used to test for association with quantitative lipid traits including plasma total cholesterol (TC), LDL cholesterol (LDL-C), HDL cholesterol (HDL-C), non-HDL-C and triglyceride concentration (TG) within each Inuit population as well as apolipoprotein B (APOB) concentration where available. The four Inuit populations were then combined and re-assessed for association with lipid traits. The *LDLR* variants were genotyped using either TaqMan SNP genotyping assays (Applied Biosystems; Foster City CA) or direct Sanger sequencing. Apolipoprotein E (APOE) isoforms were inferred based on haplotypes using

Figure 2.2 Structural organization of the human LDL receptor protein and the relative positions of the p.G116S and p.R730W variants. The 839-amino acid mature protein is shown here with corresponding exon and domain annotations. Modified from Hobbs *et al* (1990).

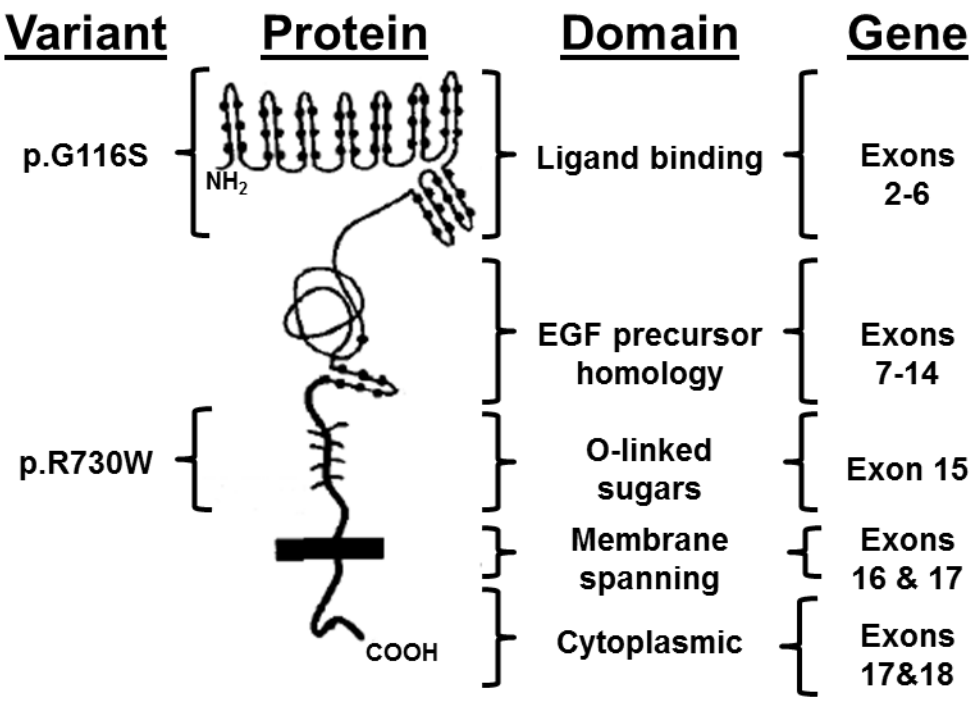

Variant
Protein
Domain
Gene
p.G116S
NH2
Ligand binding
Exons 2-6
EGF precursor homology
Exons 7-14
p.R730W
O-linked sugars
Exon 15
Membrane spanning
Exons 16 & 17
Cytoplasmic
Exons 17&18
COOH

Table 2.2. *In silico* analyses of p.G116S and p.R730W on LDLR function.

Position	Variant	Nucleotide substitution	*in silico* algorithm	Score	Prediction
Chr19:11,215,991	p.G116S Exon 4	c.409G>**A**	PMUT	n.d.	Neutral
			SIFT	0.01	Damaging
			PolyPhen	0.999	Probably damaging
			MutPred	n.d.	0.81 probability of deleterious mutation
Chr19:11,233,960	p.R730W Exon 15	c. 2251C>**T**	PMUT	n.d.	Pathological
			SIFT	0.04	Damaging
			PolyPhen	0.951	Possibly damaging
			MutPred	n.d.	0.51 probability of deleterious mutation

Amino acid positions for mutations refer to the mature protein. Abbreviations as in Table 2.1.

rs429358 and rs7412 which encode amino acid substitutions p.C112R and p.C158R respectively (Fullerton et al., 2000). *APOE* genotypes for SNPs rs429358 and rs7412 were also determined using pre-designed TaqMan assays.

2.2.3 Statistical analysis

Study cohort demographics were evaluated within each Inuit subpopulation using t-tests or ANOVA for continuous variables using SAS v9.2 (Cary, NC); statistical significance was defined as $P<0.05$. We tested for association between *LDLR* variant genotypes and lipid-related traits using multivariate linear regression within each Inuit subpopulation using an additive genetic model adjusted for age, sex and BMI as previously reported (Lanktree et al., 2009). Multi-variate regression in the combined Inuit cohort, comprising all 5 Inuit population samples, was additionally adjusted for geographic location. Regression analyses between *LDLR* genotypes and plasma lipid traits, pairwise linkage disequilibrium and haplotype phase were investigated using the PLINK bioinformatics toolkit (http://pngu.mgh.harvard.edu/purcell/plink/) (Purcell et al., 2007).

2.2.4 Bioinformatic analysis

Variant effects on LDLR function were predicted using PMUT (Ferrer-Costa et al., 2005), Polyphen (Adzhubei et al., 2010), MutPred (Li et al., 2009) and SIFT (Ng and Henikoff, 2001) variant modeling algorithms. Algorithm scores for p.G116S and p.R730W were included where available. Evolutionary conservation was investigated across species at amino acid positions in the vicinity of p.G116S and p.R730W using the BLAST alignment tool which aligns homologous regions from a range of available species (Kent, 2002). Reported FH mutations in *LDLR* were referenced from the Human

Gene Mutation Database (Stenson et al., 2009). The 1000 Genomes Project and the Exome Variant Server (http://evs.gs.washington.edu/EVS/) variant databases were also referenced for previous reports of p.G116 or p.R730W (Abecasis et al., 2012). The 1000 Genomes Project reports variants with frequencies >1% from 1,092 sequenced genomes from multiple ethnicities. The Exome Variant Server reports variants from the National Heart, Lung, and Blood Institute (NHLBI) GO Exome Sequencing project which maintains exome sequencing data on more than 200,000 individuals.

2.3 RESULTS

2.3.1 Study subjects

Demographic data from five Inuit cohorts are shown in **Table 2.1**. Overall, demographic attributes between Inuit cohorts were comparable with each sample population represented by average ages >35 years and female participants representing the majority or >50% of the sample. Average plasma lipid traits were also consistent between Inuit populations.

2.3.2 *LDLR* variant discovery and frequencies

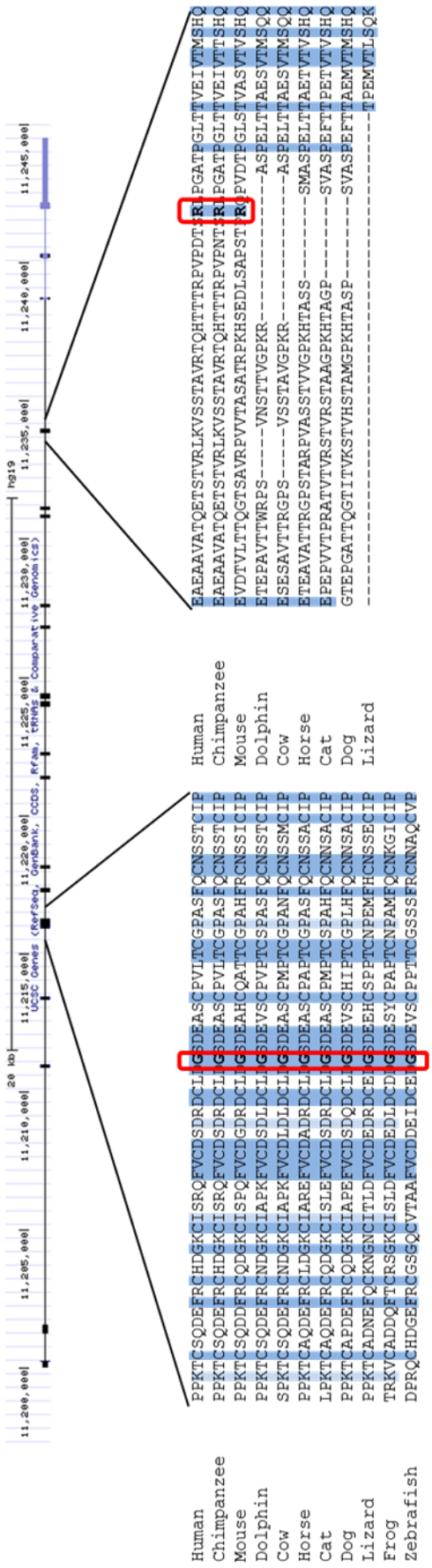
First, p.G116S and p.R730W were detected following direct Sanger sequencing of *LDLR* exons within a subgroup of the Greenland cohort with LDL-C >6.00 mmol/L (n=10). The Greenland population sample was selected for Sanger sequencing as it was the only sample available at the initiation of this study. We then sought to establish p.G116S and p.R730W variant frequencies within the five additional Inuit cohorts and a combined cohort incorporating all five Inuit sample populations. Both variants were observed with

common frequencies >1% within each Inuit population (**Table 2.1**). The p.G116S variant ranged in frequency from 2% in the Keewatin cohort to 13% in the Greenland cohort with an overall frequency of 10% when all cohorts were combined. The p.R730W variant was consistently observed at a high frequency ranging from 11% in the Greenland cohort to 17% in the Keewatin cohort with a combined frequency of 14%. The p.G116S and p.R730W were both absent from an indigenous Canadian First Nations population sample. Furthermore, neither G116S nor R730W had been reported by the 1000 Genomes Project or the NHLBI GO Exome Sequencing Project.

2.3.3 *In silico* analyses suggest p.G116S and p.R730W introduce deleterious effects on LDLR function

As the functional domains of LDLR have been well-characterized (**Figure 2.2**), we determined that the p.G116S variant in exon 4 was localized within the ligand binding domain. Out of the 1,763 hypercholesterolemia mutations identified in *LDLR* to date, 183 are found within exon 4 suggesting that this locus is a hot-spot for FH-related mutation (Stenson et al., 2009). Using BLAST sequence alignment, we observed that glycine at the 116th amino acid in LDLR was conserved across 10 additional orthologous LDLR homologs further suggesting that mutations at this amino acid are not well tolerated (**Figure 2.3**). A glycine-to-serine substitution introduces a polar residue of higher molecular weight and greater hydrophilicity which may impact upon local folding particularly as exon 4 encodes 3 of 7 cystein-rich repeats found within the ligand binding domain with each repeat modulating ligand binding (Hobbs et al., 1990). Accordingly,

Figure 2.3 Amino acid conservation in the vicinity of p.G116S and p.R730W. Multiple amino acid residue sequence alignments from divergent species show conservation at amino acids 116 and 730 in *LDLR* (highlighted in red). Amino acid residues conserved between homologs are highlighted in blue and indicate local conservation.


hg19
20 kb
11,200,000
11,205,000
11,210,000
11,215,000
11,220,000
11,225,000
11,230,000
11,235,000
11,240,000
11,245,000
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
Human
Chimpanzee
Mouse
Dolphin
Cow
Horse
Cat
Dog
Lizard
Frog
Zebrafish

we investigated the predicted effects of p.G116S on LDLR function using four independent mutation prediction algorithms (**Table 2.2**). Three algorithms predicted damaging effects of G116S on LDLR function while the PMUT algorithm predicted a neutral effect. Another variant at p.G116, a glycine-to-cysteine substitution (p.G116C), was previously reported in Polish FH patient and was predicted to have a deleterious effect on LDLR function (Chmara et al., 2010). The report of p.G116C provides additional support that G116 is an important residue in LDLR function.

Similar bioinformatic analyses were performed using p.R730W. p.R730W is located in exon 15 which encodes an attachment site for O-linked carbohydrate chains and has no clear functional role in LDLR activity (Hobbs et al., 1990). In comparison to exon 4, only 19 hypercholesterolemia mutations have been reported in exon 15 (Stenson et al., 2009). Sequence conservation in the vicinity of p.R730W is also comparatively less strict suggesting mutations within this exon may be more tolerable than in exon 4 (**Figure 2.3**). An arginine-to-tryptophan substitution introduces a larger molecular weight residue with a shift from polar to neutral charge and decreased hydrophilicity; however it is not clear how this substitution may affect the binding of O-linked sugars at this domain or impact LDLR function. Mutation prediction algorithms all predicted deleterious effects on LDLR function for p.R730W. However, milder effects were predicted relative to p.G116S as MutPred predicted a lower probability of a deleterious effect compared to p.G116S and Polyphen reported p.R730W as “Possibly damaging” as opposed to “Probably damaging”. At the p.R730W residue, an arginine-to-glutamine (p.R730Q) mutation was reported in a Dutch FH cohort and was predicted as “Tolerated” and “Low

Risk" (Fouchier et al., 2005). Although p.R730Q was identified in a cohort of FH patients, the frequency of the mutation is not known nor is it clear whether carriers of p.R730Q may have carried additional FH-causing mutations. Thus the importance of residue arginine at amino acid 730 does not appear as robust as glycine at amino acid 116 in maintaining LDLR function.

2.3.4 Mean lipid traits differ based on p.G116S or p.R730W genotype

As LDLR is a major regulator of plasma cholesterol homeostasis, we tested for association between both p.G116S or p.R730W carrier status and plasma lipid traits. Within each population sample, carriers of p.G116S had significantly higher average TC and LDL-C concentrations compared to non-carriers by ~0.7 mmol/L for p.G116S heterozygotes in the combined Inuit cohort (**Table 2.3A**); mean TC and LDL-C among p.G116S homozygotes for serine at amino acid 116 was nearly 1 mmol/L higher than homozygotes for glycine at amino acid 116. Mean apoB and non-HDL-C concentrations were also consistently higher among p.G116S carriers compared to homozygotes for glycine at amino acid 116 (**Table 2.3A, 2.3B**). Conversely, p.R730W carrier status was not robustly linked with any lipid trait within the individual Inuit cohorts. However, examination of the combined Inuit cohort revealed a significantly lower mean LDL-C and non-HDL-C concentrations as well as higher HDL-C in p.R730W carriers versus non-carriers (**Table 2.3A, 2.3B)**.

Table 2.3A Mean lipid traits based on p.G116S or p.R730W genotype.

	TC			LDL-C			APOB		
G116S genotype	**GG**	**GA**	**AA**	**GG**	**GA**	**AA**	**GG**	**GA**	**AA**
Greenland	5.74±1.09	6.40±1.07	6.70±1.32**	3.63±0.99	4.34±0.98	4.70±1.12**	0.89±0.22	1.02±0.24	1.10±0.20**
Keewatin	4.93±0.97	6.02±0.97*	n.d.	3.02±0.87	4.11±0.90*	n.d.	0.95±0.24	1.36±0.12**	n.d.
Inuvik	5.00±0.98	5.58±1.07	5.05±0.11*	2.88±0.86	3.43±1.03	3.27±0.20*	0.90±0.25	1.03±0.25	0.95±0.04*
Nunavik	4.90±0.94	5.38±1.13	6.27±0.58**	2.72±0.82	3.11±0.97	3.90±0.77**	0.94±0.24	1.02±0.24	1.24±0.19*
Alaska	5.16±1.10	5.62±1.32	5.57±0.96**	3.12±0.91	3.58±1.17	3.61±0.96**	n.d.	n.d.	n.d.
Combined	5.29±1.11	5.94±1.24	6.20±1.25**	3.22±0.98	3.88±1.14	4.21±1.14**	0.90±0.23	1.03±0.24	1.11±0.20**

	TC			LDL-C			APOB		
R730W genotype	**CC**	**CT**	**TT**	**CC**	**CT**	**TT**	**CC**	**CT**	**TT**
Greenland	5.91±1.12	5.90±1.17	6.1±1.09	3.83±1.03	3.78±1.08	3.73±0.72	0.92±0.22	0.91±0.24	0.95±0.31
Keewatin	4.98±1.00	4.87±1.00	5.18±0.91	3.08±0.90	2.96±0.92	3.19±0.76	0.97±0.25	0.95±0.26	0.98±0.22
Inuvik	5.10±1.01	4.86±0.97	5.15±0.81	2.97±0.90	2.78±0.86	2.98±0.40	0.92±0.26	0.87±0.24	0.91±0.18
Nunavik	4.94±0.96	4.93±0.90	5.52±0.93	2.80±0.84	2.71±0.85	3.14±1.04	0.95±0.23	0.93±0.24	1.04±0.28
Alaska	5.23±1.16	5.29±1.14	5.18±1.08	3.21±1.00	3.22±0.94	3.19±0.85	n.d.	n.d.	n.d.
Combined	5.43±1.17	5.37±0.17	5.47±1.07	3.37±1.06	3.28±1.04	3.20±0.82*	0.93±0.23	0.91±0.24	0.97±0.27

* indicates P<0.05 and **indicates P<0.0001 using ANOVA adjusted for age, sex and BMI. Abbreviations as in Table 2.1.

Table 2.3B Mean lipid traits based on p.G116S or p.R730W genotype.

	HDL-C			Non-HDL-C			TG		
G116S genotype	**GG**	**GA**	**AA**	**GG**	**GA**	**AA**	**GG**	**GA**	**AA**
Greenland	1.58±0.45	1.54±0.40	1.58±0.48	4.16±1.09	4.84±1.12	5.13±1.18**	1.18±0.68	1.16±0.66	0.96±0.30
Keewatin	1.44±0.41	1.51±0.47	n.d.	3.48±0.94	4.51±0.89*	n.d.	1.01±0.55	0.87±0.39	n.d.
Inuvik	1.35±0.41	1.39±0.40	1.17±0.15	3.65±1.03	4.20±1.03	3.88±0.04*	1.74±1.32	1.86±0.81	1.34±0.54
Nunavik	1.61±0.42	1.71±0.47	1.92±0.77	3.25±0.99	3.68±1.05	4.35±0.90*	1.24±0.75	1.25±0.56	0.99±0.32
Alaska	1.64±0.44	1.64±0.45	1.56±0.46	3.52±1.03	3.98±1.26	4.01±1.01**	0.96±0.58	0.90±0.47	0.94±0.47
Combined	1.57±0.44	1.59±0.43	1.58±0.49	3.71±1.09	4.35±1.24	4.62±1.20**	1.13±0.75	1.09±0.62	0.97±0.38

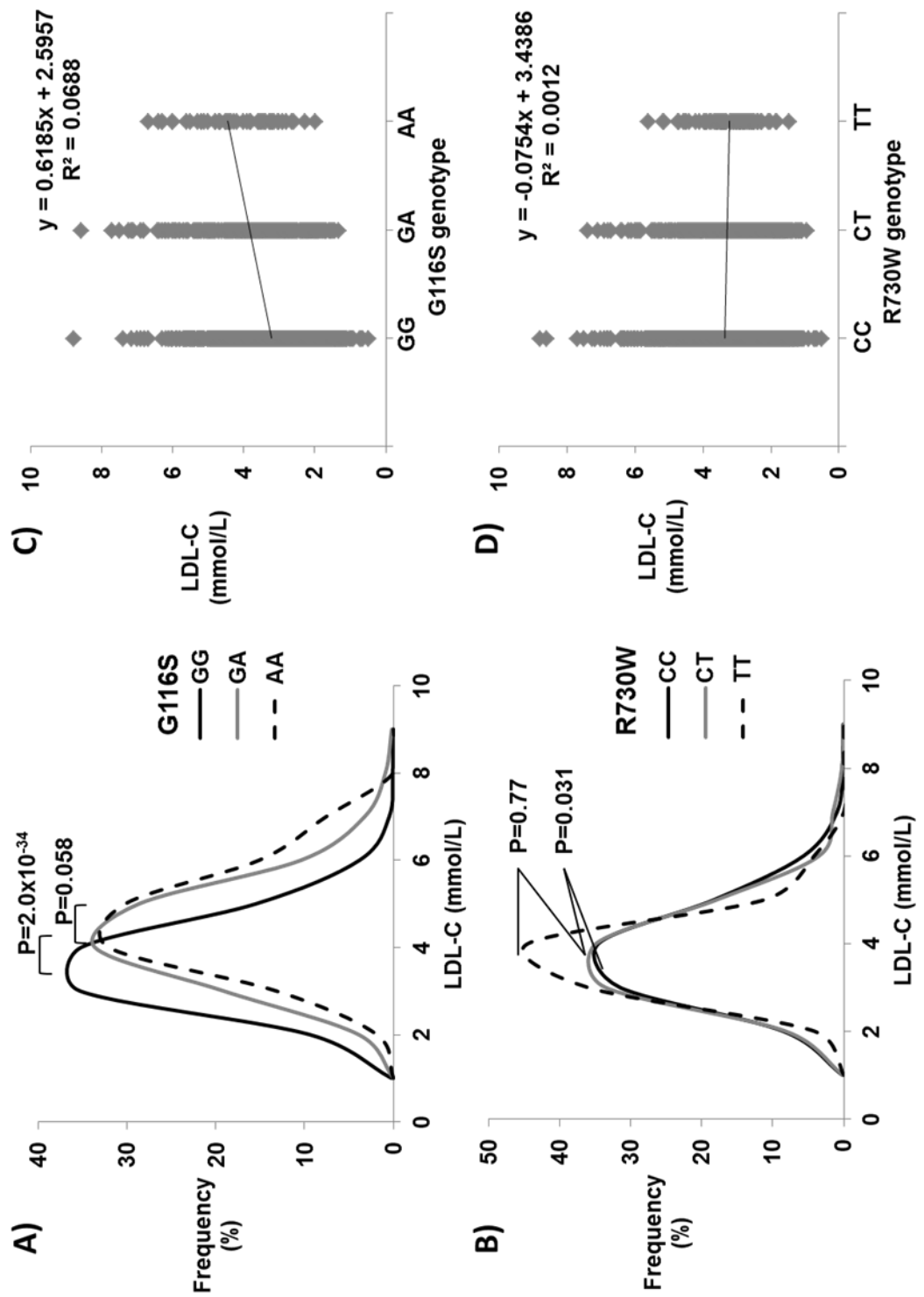
	HDL-C			Non-HDL-C			TG		
R730W genotype	**CC**	**CT**	**TT**	**CC**	**CT**	**TT**	**CC**	**CT**	**TT**
Greenland	1.57±0.44	1.56±0.45	1.65±0.46	4.33±1.13	4.34±1.20	4.45±1.17	1.14±0.64	1.24±0.75	1.35±0.95
Keewatin	1.43±0.41	1.46±0.37	1.65±0.53	3.57±0.97	3.41±1.02	3.53±0.80	1.04±0.54	0.98±0.59	0.74±0.28
Inuvik	1.36±0.41	1.40±0.46	1.42±0.35	3.73±1.05	3.48±0.99	3.67±0.76	1.80±1.38	1.54±0.78	1.70±0.72
Nunavik	1.61±0.43	1.67±0.42	1.83±0.38	3.31±0.98	3.23±0.99	3.69±1.17	1.56±0.63	1.21±0.72	1.20±0.44
Alaska	1.62±0.43	1.68±0.47	1.70±0.42	3.62±1.10	3.61±1.05	3.48±1.06	0.95±0.58	0.95±0.52	1.00±0.54
Combined	1.56±0.44	1.61±0.46	1.68±0.45*	3.86±1.15	3.76±0.16	3.80±1.10*	1.12±0.76	1.11±0.67	1.11±0.68

* indicates P<0.05 and **indicates P<0.0001 using ANOVA adjusted for age, sex and BMI. Abbreviations as in Table 2.1.

2.3.5 p.G116S is associated with LDL-C concentration

As LDL-C appeared to underlie the differences in cholesterol concentrations in S116 and W730 carriers, we next sought to investigate the effects of G116S and R730W on LDL-C using genetic models adjusted for age, sex and BMI. Frequency distributions based on p.G116S within the combined Inuit cohort suggested an additive effect on LDL-C concentration. Mean LDL-C concentrations were significantly different between homozygotes for glycine at amino acid 116 and p.G116S heterozygotes ($P=2.0x10^{-34}$, **Figure 2.4A**). Furthermore, the difference in mean LDL-C concentrations between p.G116S heterozygotes and homozygotes approached significance (P=0.058, **Figure 2.4A**). p.R730W genotype did not follow a distinct genetic model as the mean LDL-C for homozygotes for arginine at amino acid 730 differed significantly from the mean LDL-C concentration of p.R730W heterozygotes (P=0.031); however, mean LDL-C did not differ significantly between p.R730W heterozygotes and homozygotes for tryptophan at amino acid 730 (P=0.77, **Figure 2.4B**). Plotting LDL-C concentrations for all participants based on p.G116S or p.R730W genotype revealed linear trends between LDL-C and each additional copy of either the p.G116S or p.R730W variant (**Figure 2.4C, 2.4D**). We used multi-variate linear regression to test whether the observed variant-LDL-C trends also fit linear models. Within each Inuit cohort, p.G116S was associated with increased LDL-C. In a combined cohort of all Inuit population samples, each copy of the S116 variant was associated with a ~0.54 mmol/L increase in LDL-C ($P=5.6x10^{-49}$, **Table 2.4**). W730 was non-significantly linked with lower LDL-C within each Inuit cohort and was linked with an overall lowering effect on LDL-C by ~0.05 mmol/L (P=0.13, **Table 2.4**). p.G116S and p.R730W variants were not in strong linkage disequilibrium ($r^2=0.017$) and were not

Figure 2.4 *LDLR* variants and trends with LDL-C in a combined Inuit cohort. Inuit participants were separated based on p.G116S or p.R730W genotype and LDL-C concentration. **A)** Frequency distribution of Inuit participants based on p.G116S carrier status and LDL-C concentration. Mean LDL-C concentrations between non-carriers (GG, n=2585), heterozygotes (GA, n=559) and homozygotes (AA, n=53) were compared using t-tests with P-values indicated. **B)** Frequency distributions were similarly constructed for p.R730W non-carriers (CC, n=2408), heterozygotes (CT, n=717) and homozygotes (TT, n=72) and were also compared using t-tests. **C)** Distribution of LDL-C concentrations based on p.G116S genotype per study participant following an additive genetic model with a calculated line of best fit. **D)** Distribution of LDL-C concentrations per participant were similarly plotted for p.R730W genotype.


A)
P=2.0x10-34
P=0.058
G116S
GG
GA
AA
Frequency (%)
LDL-C (mmol/L)
B)
P=0.77
P=0.031
R730W
CC
CT
TT
Frequency (%)
LDL-C (mmol/L)
C)
y = 0.6185x + 2.5957
R² = 0.0688
LDL-C (mmol/L)
GG
GA
AA
G116S genotype
D)
y = -0.0754x + 3.4386
R² = 0.0012
LDL-C (mmol/L)
CC
CT
TT
R730W genotype

predicted to be in phase within the same haplotype. We therefore could not investigate the effect of a haplotype containing both p.G116S and p.R730W on any lipid traits.

2.3.6 p.G116S effect on LDL-C is greater than APOE E4 and common LDL-C GWAS variants

The p.G116S variant represents a unique combination of both high variant frequency (~10%, **Table 2.1**) and large effect size (~0.54 mmol/L per p.G116S allele, **Table 2.4**). APOE E4 isoform has similarly been established as a high-frequency variant robustly associated with LDL-C (Khan et al., 2013; Ward et al., 2009). We therefore tested for association between APOE E4 and LDL-C in the Inuit population samples in order to compare effect sizes between the novel p.G116S variant and the established APOE E4 isoform. APOE E4 frequencies were comparable across the four Inuit populations in which APOE isoform status was available and ranged from 21% to 27% (**Table 2.5**). Using multi-variate linear regression to estimate the per-allele effect size on LDL-C, we identified a robust association between APOE E4 and LDL-C exclusively within the Greenland population sample (0.22 mmol/L, $P=8.2x10^{-6}$; **Table 2.5**); in a combined cohort of all Inuit population samples the effect was diminished but remained significant (0.15 mmol/L, $P=1.8x10^{-5}$; **Table 2.5**). In comparison, the p.G116S-associated effect on LDL-C was almost 4-fold greater than APOE E4 in combined Inuit cohorts.

GWAS have comprehensively identified robust associations between common variants and LDL-C (Teslovich et al., 2010). In order to further give context to the p.G116S-associated frequency and effect on LDL-C, we compared p.G116S to the most

Table 2.4 Associations between two *LDLR* variants and LDL-C.

Variant	Population	β (mmol/L)	SE	P-value
p.G116S	Greenland	0.64	0.05	$1.8x10^{-30}$
	Keewatin	1.02	0.27	$1.7x10^{-4}$
	Inuvik	0.52	0.16	0.0011
	Nunavik	0.40	0.09	$3.7x10^{-5}$
	Alaska	0.41	0.06	$9.4x10^{-12}$
	Combined	0.54	0.04	$5.6x10^{-49}$
p.R730W	Greenland	-0.11	0.06	0.077
	Keewatin	-0.003	0.09	0.97
	Inuvik	-0.13	0.11	0.24
	Nunavik	-0.05	0.08	0.52
	Alaska	-0.01	0.05	0.85
	Combined	-0.05	0.03	0.13

Effect sizes and P-values are based on the minor alleles p.G116S or p.R730W. SE, standard error. Greenland (n=1162) Keewatin (n=204) Inuvik (n=253) Nunavik (n=389) Alaska (n=1113) Combined (n=3121).

Table 2.5 APOE E4 effect on LDL-C in select Inuit populations.

Population	E4 frequency	β (mmol/L)	SE	P-value
Greenland	0.22	0.22	0.049	$8.2x10^{-6}$
Keewatin	0.21	0.06	0.10	0.56
Inuvik	0.23	0.006	0.10	0.95
Nunavik	0.27	0.09	0.06	0.17
Combined	0.23	0.15	0.03	$1.8x10^{-5}$

ApoE E4 effect sizes were calculated in comparison to E3 carriers using linear regression adjusted for age, sex and BMI in Keewatin (n=200), Greenland (n=1096), Inuvik (n=212), Nunavik (n=383), and a combined cohort (n=1891). ApoE E4 frequencies were calculated from larger populations including ApoE E2, E3 and E4 carriers.

significant LDL-C GWAS variants which we listed in **Table 2.6.** Together, the strongest association signal at *SORT1* as well as associations at candidate LDL-C loci such as *LDLR*, *APOB* and *PCSK9* corresponded to effect sizes ranging from 0.05 mmol/L to 0.18 mmol/L which are overshadowed by the ~0.54 mmol/L per-allele effect size we ascribed to p.G116S.

2.3.7 Mean IMT is not linked with p.G116S or p.R730W genotype

Mean IMT is an established marker of atherosclerosis and is correlated with plasma LDL-C concentrations (Negi and Nambi, 2012; Sun et al., 2000). As we identified associations between p.G116S and p.R730W, and LDL-C, we sought to test whether either *LDLR* variant genotype was also linked with changes in IMT. Using available IMT measurements from the Inuvik and Nunavik population samples, we compared mean IMT measurements between the *LDLR* variant genotypes and observed no significant difference suggesting no clear effect on IMT based on *LDLR* variant genotype (**Table 2.7**).

2.4 DISCUSSION

The major finding of our study is the discovery of two common *LDLR* variants, p.G116S and p.R730W, which are private among Inuit populations. Furthermore, we showed that the p.G116S variant was robustly associated with a large increase in plasma LDL-C while p.R730W showed a modest non-significant LDL-C-lowering effect. The p.G116S variant was also unique due to the high frequency of the variant coupled with a large effect size

Table 2.6 The most significant LDL-C-associated common variants.

CHR	Locus	SNP	β (mmol/L)	P-value
1	*SORT1*	rs629301	0.15	$1x10^{-170}$
19	*APOE*	rs4420638	0.18	$9x10^{-147}$
19	*LDLR*	rs6511720	0.18	$4x10^{-117}$
2	*APOB*	rs1367117	0.10	$4x10^{-114}$
2	*ABCG5/8*	rs4299376	0.07	$2x10^{-47}$
1	*PCSK9*	rs2479409	0.05	$2x10^{-28}$

Effect sizes and P-values were reported by Teslovich *et al.* (Teslovich et al., 2010)

Table 2.7 IMT measurements based on p.G116S and p.R730W genotypes.

	C-IMT			Average IMT		
p.G116S genotype	**GG**	**GA**	**AA**	**GG**	**GA**	**AA**
Inuvik	0.73±0.17 (n=54)	0.73±0.20 (n=7)	n.d.	0.56±0.16 (n=41)	0.50±0.21 (n=6)	n.d.
Nunavik	0.78±0.16 (n=83)	0.74±0.13 (n=20)	n.d.	0.83±0.19 (n=5)	0.77±0.20 (n=4)	n.d.
p.R730W genotype	**CC**	**CT**	**TT**	**CC**	**CT**	**TT**
Inuvik	0.73±0.15 (n=44)	0.74±0.19 (n=15)	0.59±0.03 (n=2)	0.54±0.17 (n=34)	0.59±0.15 (n=12)	0.57±0.00 (n=1)
Nunavik	0.77±0.16 (n=74)	0.75±0.16 (n=24)	0.86±0.14 (n=5)	0.84±0.19 (n=7)	0.68±0.01 (n=2)	n.d.

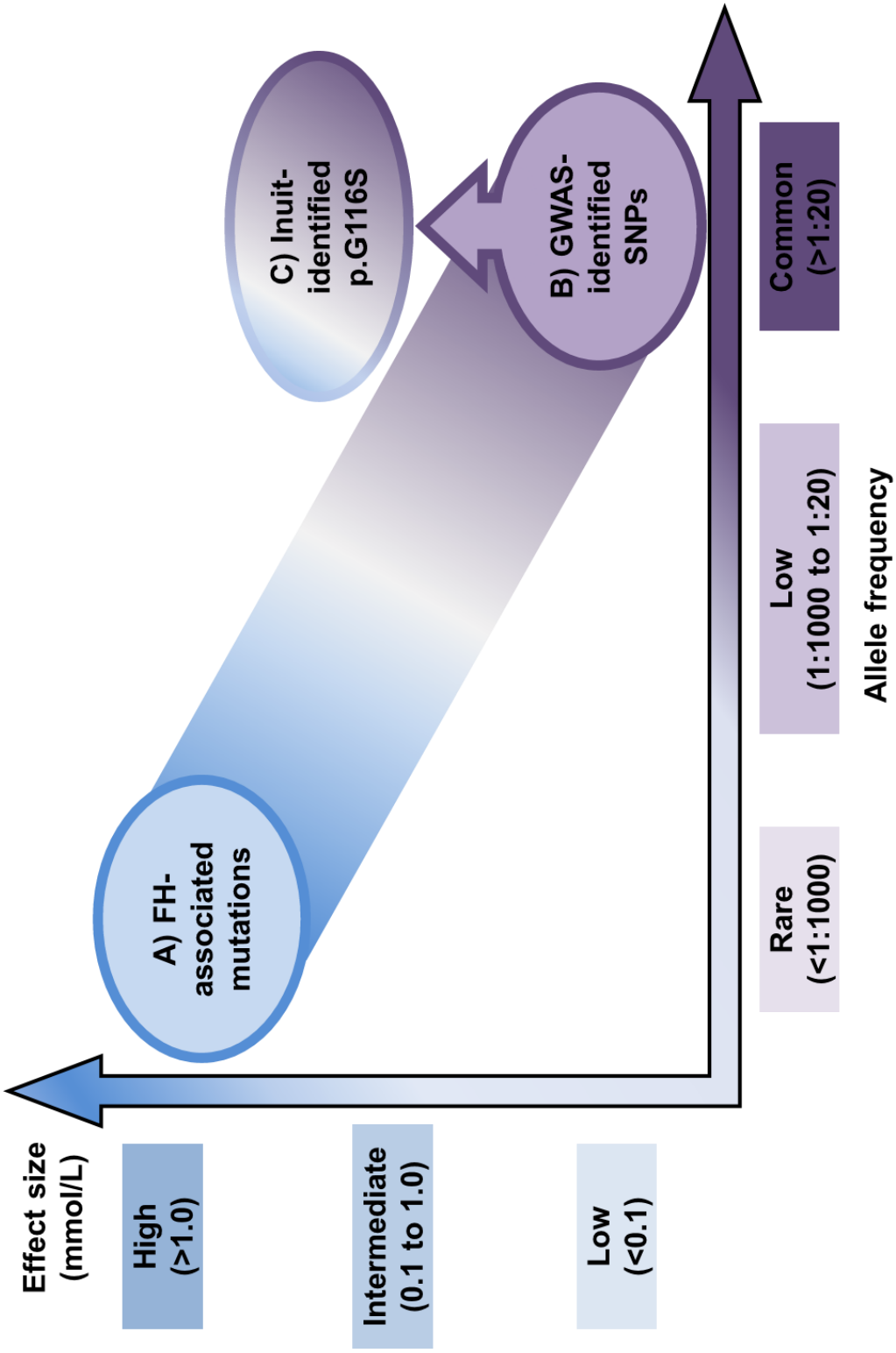
Statistical significance was tested using ANOVA. Inuvik participants with C-IMT (n=61) and average IMT (n=47), and Nunavik participants with C-IMT (n=103) and average IMT (n=9) were included for analysis. Abbreviations as in Table 2.1; C-IMT, common carotid intima-media thickness. All units are expressed in mm.

on LDL-C. Our study has thus proposed a genetic CVD risk factor exclusive to Inuit descendants with potential clinical utility.

Our discovery of the association between the p.G116S variant and LDL-C concentration is of particular interest from a public health perspective as circumpolar Inuit communities are currently facing increased risk of CVD compared to non-Inuit populations. Early Inuit population studies propagated the concept of low CVD mortality among Inuit communities based on marine diet and a possible genetic component, however, subsequent analysis have largely dispelled this myth (Bjerregaard et al., 2003b). As introduced earlier, subsequent studies on CVD in Inuit communities have suggested greater CVD risk within Inuit communities compared to non-Inuit populations as Westernization influences the lifestyles of the younger generation. Given that every 1 mmol/L increase in LDL-C corresponds to a ~21% increase in CVD and ~16% increase in all-cause mortality, the 0.54 mmol/L increase in LDL-C per p.G116S allele could potentially lead to ~10% increases in CVD and all-cause mortality respectively (Gould et al., 2007).

In addition to the implications on LDL metabolism, the p.G116S and p.R730W variants are distinct based on the high frequencies observed within our study cohorts. Within the context of the common disease-common variant hypothesis (CDCV), which proposed that a limited number of common variants underlies common complex disease etiology (Lander, 1996), the large effect size associated with p.G116S was unexpected given the high minor allele frequency (**Figure 2.5**). The APOE E4 isoform represents an

Figure 2.5 The common disease-common variant hypothesis in relation to the G116S variant. LDL-associated variants have adhered to the CDCV hypothesis-predicted trend where **A)** low-frequency mutations, as observed in FH, contribute large effects on LDL concentrations whereas **B)** common variants, as identified through GWAS on LDL, are typically associated with small effects on plasma LDL. The novel **C)** G116S variant is unique as it represents a common variant with a considerable effect on plasma LDL. Modified from Manolio et al. (2009).


Effect size (mmol/L)
High (>1.0)
Intermediate (0.1 to 1.0)
Low (<0.1)
A) FH-associated mutations
B) GWAS-identified SNPs
C) Inuit-identified p.G116S
Rare (<1:1000)
Low (1:1000 to 1:20)
Common (>1:20)
Allele frequency

established high-frequency variant with a robust association with higher LDL-C. Compared to E3 homozygotes, each copy of E4 was associated with a 0.16 mmol/L increase in LDL-C (Khan et al., 2013). We replicated a similar effect size in a combined Inuit cohort (0.15 mmol/L per copy of E4; **Table 2.5**); however, it was clear that S116 had a considerably greater impact on LDL-C. As the top LDL-C SNPs from across the genome reflected similar effect sizes as APOE E4, which ranged from 0.05 mmol/L to 0.18 mmol/L **(Table 2.6**), it is clear that the combination of the G116S frequency and effect size is anomalous. Despite the large effect size of p.G116S, it is interesting that p.G116S homozygotes do not appear to express the dramatic phenotypes that are observed with patients homozygous for FH-causing mutations.

The LDLR amino acid positions of 116 and 730 provide insight on potential effects on LDLR function (**Figure 2.2**). Amino acid 116 lies within exon 4 of *LDLR* which encodes the ligand binding domain. As this domain is important for the binding and internalization of apoB-containing lipoprotein particles, p.G116S can potentially perturb binding affinity and thus LDLR activity; this hypothesis was supported by predictions of potentially damaging effect by two *in silico* algorithms. Exon 15, which contains amino acid 730, is enriched for serine and threonine residues which facilitate attachment of O-linked carbohydrate chains; however, the absence of exon 15 has not been associated with any significant functional consequence *in vitro* (Hobbs et al., 1990). This observation runs contrary to both our *in silico* analysis which anticipated potentially damaging effects on LDLR function as well as the potential gain-of-function effect that we observed with p.R730W in modest lowering of LDL-C. In order to better understand

the effects of either variant on LDLR function, direct biochemical analyses measuring LDLR expression and activity must be implemented.

Our findings also provide further support for the concept of a unique genomic architecture within the Inuit. It remains a question as to how p.G116S and p.R730W reached high frequency across the circumpolar Inuit populations; however, this phenomenon is likely explained by the founder effect. Various sources of evidence have supported the original founding human populations in the Americas by Asians (Hey, 2005). It is possible that carriers of p.G116S and p.R730W were among the small founder populations that migrated eastward from Alaska towards Greenland. As founder populations in geographic isolation are able to expand in numbers, limited genetic heterogeneity facilitates the inflation of allele frequencies that may have been rare in a larger non-related population. Similar founder effects have historically been observed, particularly in the case of FH frequency where founder populations such as the Quebecois in Canada and Dutch immigrants to South Africa report remarkably high FH frequencies due to the propagation of founder mutations (Liyanage et al., 2011). Through mechanisms such as the founder effect, it is therefore possible for variants to gain high frequency despite potentially negative effects on health and mortality particularly as HeFH mutations are less penetrant and do not increase selective pressure before reproductive age.

In addition to the p.G116S and p.R730W variants described here, the p.P479L substitution in the carnitine palmitoyltransferase IA gene (*CPT1A*) has been identified as

a private variant among Inuit and Canadian First Nations populations where the p.P479L variant was associated with hypoketosis and hypoglycemia (Brown et al., 2001; Lemas et al., 2012; Rajakumar et al., 2009). Together, the identification of these high-frequency variants with large effect sizes suggests that Inuit descendants may possess a unique genetic architecture with effects on cardio-metabolic traits that are not fully identified nor understood. With the increasing viability of whole genome and exome sequencing, it will be possible to perform a comprehensive scan for genetic variation within the Inuit. Additional evidence of unique variation relating to cardiovascular health will further support the concept that Inuit communities may be exposed to unique CVD risk which may require unique guidelines for more efficient and targeted CVD prevention and management.

A potential limitation within this study pertains to our limited information on participant relatedness. As Inuit communities are generally isolated with low net population migration, it is expected that genetic heterogeneity is lower compared to the level that may be observed in the general population. For the purpose of genetic association, limited genetic heterogeneity is considered advantageous as this limits the pool of variants present within the population and thus limits the probability of false positive association. Conversely, closely related individuals may have a similar lipid profile due to shared environmental or additional genetic factors which may contribute to spurious associations. Future studies will require detailed family structure and relatedness data in order to adjust for the effects of relatedness on observed associations.

Furthermore, our claim that p.G116S is private to Inuit descendants was questioned by the previous report of p.G116S in a Danish FH cohort (Damgaard et al., 2005). As Greenland continues to share a history of migration and sociocultural interaction with Denmark, it is possible that a patient or patients with hypercholesterolemia as well as Inuit ancestry were included in the Danish FH cohort. However, this remains speculative as records on patient ethnicity were not published.

Our study was also limited to reporting on the discovery of p.G116S and p.R730W, the associations with plasma LDL-C concentration and *in silico* prediction analysis. We sought to test for association with IMT; an established marker of CVD. However, no robust association was detected between IMT and *LDLR* variant genotype. While this experiment suggested that the 0.54 mmol/L increase in LDL-C associated with p.G116S does not correlate with thickening of the carotid artery walls, a major caveat to this interpretation was the limited sample size as well as variability in the application of IMT measurement methodology. Functional studies of the potential effects of these variants on LDLR bioavailability remain to be performed but are crucial in elucidating the mechanism underlying the robust association between p.G116S and LDL-C concentration and establishing a causal effect between p.G116S and elevated LDL-C. Going forward, we propose to first assess LDLR expression based on G116S and R730W genotype within *in vitro* Chinese hamster ovary cell-based models and immunoblotting followed by assessment of LDLR activity via fluorescently-labelled LDL uptake assays.

In summary, we have discovered the presence of two common *LDLR* variants that are exclusive among Inuit descendants from five distinct communities within Alaska, Canada and Greenland. We have further identified a strong association between p.G116S and plasma LDL-C concentration which has implicated p.G116S in CVD risk. The clinical utility of these findings in assessing CVD risk prediction is not presently clear as robust statistics on Inuit CVD mortality are not currently available and that Inuit communities continue to undergo progressive westernization which may affect CVD risk and prevalence over the coming decades. Additional studies involving biological assessment of p.G116S and p.R730W on LDLR function as well as broader investigation in outstanding circumpolar Inuit populations will provide a greater understanding of the role played by these variants in LDL metabolism and overall CVD susceptibility.

2.5 REFERENCES

Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP), Seattle, WA, (http://evs.gs.washington.edu/EVS/: http://evs.gs.washington.edu/EVS/) [March 1, 2013 accessed].

Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature 491*, 56-65.

Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. *Nat Methods 7*, 248-249.

Bjerregaard, P., Curtis, T., Borch-Johnsen, K., Mulvad, G., Becker, U., Andersen, S., and Backer, V. (2003a). Inuit health in Greenland: a population survey of life style and disease in Greenland and among Inuit living in Denmark. *Int J Circumpolar Health 62 Suppl 1*, 3-79.

Bjerregaard, P., and Dyerberg, J. (1988). Mortality from ischaemic heart disease and cerebrovascular disease in Greenland. *Int J Epidemiol 17*, 514-519.

Bjerregaard, P., Jorgensen, M.E., and Borch-Johnsen, K. (2004). Serum lipids of Greenland Inuit in relation to Inuit genetic heritage, westernisation and migration. *Atherosclerosis 174*, 391-398.

Bjerregaard, P., Mulvad, G., and Pedersen, H.S. (1997). Cardiovascular risk factors in Inuit of Greenland. *Int J Epidemiol 26*, 1182-1190.

Bjerregaard, P., Young, T.K., and Hegele, R.A. (2003b). Low incidence of cardiovascular disease among the Inuit--what is the evidence? *Atherosclerosis 166*, 351-357.

Boyer, B.B., Mohatt, G.V., Plaetke, R., Herron, J., Stanhope, K.L., Stephensen, C., and Havel, P.J. (2007). Metabolic syndrome in Yup'ik Eskimos: the Center for Alaska Native Health Research (CANHR) Study. *Obesity (Silver Spring) 15*, 2535-2540.

Brown, N.F., Mullur, R.S., Subramanian, I., Esser, V., Bennett, M.J., Saudubray, J.M., Feigenbaum, A.S., Kobari, J.A., Macleod, P.M., McGarry, J.D., *et al.* (2001). Molecular characterization of L-CPT I deficiency in six patients: insights into function of the native enzyme. *J Lipid Res 42*, 1134-1142.

Carroll, M.D., Kit, B.K., Lacher, D.A., Shero, S.T., and Mussolino, M.E. (2012). Trends in lipids and lipoproteins in US adults, 1988-2010. *JAMA 308*, 1545-1554.

Chateau-Degat, M.L., Dewailly, E., Louchini, R., Counil, E., Noel, M., Ferland, A., Lucas, M., Valera, B., Ekoe, J.M., Ladouceur, R., *et al.* (2010). Cardiovascular burden and related risk factors among Nunavik (Quebec) Inuit: insights from baseline findings in the circumpolar Inuit health in transition cohort study. *Can J Cardiol 26*, 190-196.

Chmara, M., Wasag, B., Zuk, M., Kubalska, J., Wegrzyn, A., Bednarska-Makaruk, M., Pronicka, E., Wehr, H., Defesche, J.C., Rynkiewicz, A., *et al.* (2010). Molecular characterization of Polish patients with familial hypercholesterolemia: novel and recurrent LDLR mutations. *J Appl Genet 51*, 95-106.

Counil, E., Julien, P., Lamarche, B., Chateau-Degat, M.L., Ferland, A., and Dewailly, E. (2009). Association between trans-fatty acids in erythrocytes and pro-atherogenic lipid profiles among Canadian Inuit of Nunavik: possible influences of sex and age. *Br J Nutr 102*, 766-776.

Damgaard, D., Larsen, M.L., Nissen, P.H., Jensen, J.M., Jensen, H.K., Soerensen, V.R., Jensen, L.G., and Faergeman, O. (2005). The relationship of molecular genetic to clinical diagnosis of familial hypercholesterolemia in a Danish population. *Atherosclerosis 180*, 155-160.

Dewailly, E., Blanchet, C., Lemieux, S., Sauve, L., Gingras, S., Ayotte, P., and Holub, B.J. (2001). n-3 Fatty acids and cardiovascular disease risk factors among the Inuit of Nunavik. *Am J Clin Nutr 74*, 464-473.

Ebbesson, S.O., Adler, A.I., Risica, P.M., Ebbesson, L.O., Yeh, J.L., Go, O.T., Doolittle, W., Ehlert, G., Swenson, M., and Robbins, D.C. (2005). Cardiovascular disease and risk factors in three Alaskan Eskimo populations: the Alaska-Siberia project. *Int J Circumpolar Health 64*, 365-386.

Ebbesson, S.O., Schraer, C., Nobmann, E.D., and Ebbesson, L.O. (1996). Lipoprotein profiles in Alaskan Siberian Yupik Eskimos. *Arctic Med Res 55*, 165-173.

Ferrer-Costa, C., Gelpi, J.L., Zamakola, L., Parraga, I., de la Cruz, X., and Orozco, M. (2005). PMUT: a web-based tool for the annotation of pathological mutations on proteins. *Bioinformatics 21*, 3176-3178.

Fouchier, S.W., Kastelein, J.J., and Defesche, J.C. (2005). Update of the molecular basis of familial hypercholesterolemia in The Netherlands. *Hum Mutat 26*, 550-556.

Fullerton, S.M., Clark, A.G., Weiss, K.M., Nickerson, D.A., Taylor, S.L., Stengard, J.H., Salomaa, V., Vartiainen, E., Perola, M., Boerwinkle, E., *et al.* (2000). Apolipoprotein E variation at the sequence haplotype level: implications for the origin and maintenance of a major human polymorphism. *Am J Hum Genet 67*, 881-900.

Genest, J., McPherson, R., Frohlich, J., Anderson, T., Campbell, N., Carpentier, A., Couture, P., Dufour, R., Fodor, G., Francis, G.A., *et al.* (2009). 2009 Canadian Cardiovascular Society/Canadian guidelines for the diagnosis and treatment of dyslipidemia and prevention of cardiovascular disease in the adult - 2009 recommendations. *Can J Cardiol 25*, 567-579.

Goldstein, J.L., and Brown, M.S. (2009). The LDL receptor. *Arterioscler Thromb Vasc Biol 29*, 431-438.

Gotto, A.M., Jr., and Moon, J.E. (2012). Management of cardiovascular risk: the importance of meeting lipid targets. *Am J Cardiol 110*, 3A-14A.

Gould, A.L., Davies, G.M., Alemao, E., Yin, D.D., and Cook, J.R. (2007). Cholesterol reduction yields clinical benefits: meta-analysis including recent trials. *Clin Ther 29*, 778-794.

Gregg, E.W., Cheng, Y.J., Cadwell, B.L., Imperatore, G., Williams, D.E., Flegal, K.M., Narayan, K.M., and Williamson, D.F. (2005). Secular trends in cardiovascular disease risk factors according to body mass index in US adults. *JAMA 293*, 1868-1874.

Haase, A., and Goldberg, A.C. (2012). Identification of people with heterozygous familial hypercholesterolemia. *Curr Opin Lipidol 23*, 282-289.

Hegele, R.A. (2009). Plasma lipoproteins: genetic influences and clinical implications. *Nat Rev Genet 10*, 109-121.

Hegele, R.A., Young, T.K., and Connelly, P.W. (1997). Are Canadian Inuit at increased genetic risk for coronary heart disease? *J Mol Med (Berl) 75*, 364-370.

Hey, J. (2005). On the number of New World founders: a population genetic portrait of the peopling of the Americas. *PLoS Biol 3*, e193.

Hobbs, H.H., Russell, D.W., Brown, M.S., and Goldstein, J.L. (1990). The LDL receptor locus in familial hypercholesterolemia: mutational analysis of a membrane protein. *Annu Rev Genet 24*, 133-170.

Howard, B.V., Comuzzie, A., Devereux, R.B., Ebbesson, S.O., Fabsitz, R.R., Howard, W.J., Laston, S., MacCluer, J.W., Silverman, A., Umans, J.G., *et al.* (2010). Cardiovascular disease prevalence and its relation to risk factors in Alaska Eskimos. *Nutr Metab Cardiovasc Dis 20*, 350-358.

Jernigan, V.B., Duran, B., Ahn, D., and Winkleby, M. (2010). Changing patterns in health behaviors and risk factors related to cardiovascular disease among American Indians and Alaska Natives. *Am J Public Health 100*, 677-683.

Jorgensen, M.E., Bjerregaard, P., Kjaergaard, J.J., and Borch-Johnsen, K. (2008). High prevalence of markers of coronary heart disease among Greenland Inuit. *Atherosclerosis 196*, 772-778.

Kellett, S., Poirier, P., Dewailly, E., Sampasa, H., and Chateau-Degat, M.L. (2012). Is severe obesity a cardiovascular health concern in the Inuit population? *Am J Hum Biol 24*, 441-445.

Kent, W.J. (2002). BLAT--the BLAST-like alignment tool. *Genome Res 12*, 656-664.

Khan, T.A., Shah, T., Prieto, D., Zhang, W., Price, J., Fowkes, G.R., Cooper, J., Talmud, P.J., Humphries, S.E., Sundstrom, J., *et al.* (2013). Apolipoprotein E genotype, cardiovascular biomarkers and risk of stroke: Systematic review and meta-analysis of 14 015 stroke cases and pooled analysis of primary biomarker data from up to 60 883 individuals. *Int J Epidemiol 42*, 475-492.

Lander, E.S. (1996). The new genomics: global views of biology. *Science 274*, 536-539.

Lanktree, M.B., Anand, S.S., Yusuf, S., and Hegele, R.A. (2009). Replication of genetic associations with plasma lipoprotein traits in a multiethnic sample. *J Lipid Res 50*, 1487-1496.

Lemas, D.J., Wiener, H.W., O'Brien, D.M., Hopkins, S., Stanhope, K.L., Havel, P.J., Allison, D.B., Fernandez, J.R., Tiwari, H.K., and Boyer, B.B. (2012). Genetic polymorphisms in carnitine palmitoyltransferase 1A gene are associated with variation in body composition and fasting lipid traits in Yup'ik Eskimos. *J Lipid Res 53*, 175-184.

Li, B., Krishnan, V.G., Mort, M.E., Xin, F., Kamati, K.K., Cooper, D.N., Mooney, S.D., and Radivojac, P. (2009). Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics 25*, 2744-2750.

Liyanage, K.E., Burnett, J.R., Hooper, A.J., and van Bockxmeer, F.M. (2011). Familial hypercholesterolemia: epidemiology, Neolithic origins and modern geographic distribution. *Crit Rev Clin Lab Sci 48*, 1-18.

Middaugh, J.P. (1990). Cardiovascular deaths among Alaskan Natives, 1980-86. *Am J Public Health 80*, 282-285.

Moffatt, M.E., Young, T.K., O'Neil, J.D., Eidelheit, S., Fish, I., and Mollins, J. (1993). The Keewatin Health Assessment Study, NWT, Canada. *Arctic Med Res 52*, 18-21.

Negi, S.I., and Nambi, V. (2012). The role of carotid intimal thickness and plaque imaging in risk stratification for coronary heart disease. *Curr Atheroscler Rep 14*, 115-123.

Ng, P.C., and Henikoff, S. (2001). Predicting deleterious amino acid substitutions. *Genome Res 11*, 863-874.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., *et al.* (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet 81*, 559-575.

Rajakumar, C., Ban, M.R., Cao, H., Young, T.K., Bjerregaard, P., and Hegele, R.A. (2009). Carnitine palmitoyltransferase IA polymorphism P479L is common in Greenland Inuit and is associated with elevated plasma apolipoprotein A-I. *J Lipid Res 50*, 1223-1228.

Redwood, D.G., Lanier, A.P., Johnston, J.M., Asay, E.D., and Slattery, M.L. (2010). Chronic disease risk factors among Alaska Native and American Indian people, Alaska, 2004-2006. *Prev Chronic Dis 7*, A85.

Roy, H., Bhardwaj, S., and Yla-Herttuala, S. (2009). Molecular genetics of atherosclerosis. *Hum Genet 125*, 467-491.

Schumacher, C., Davidson, M., and Ehrsam, G. (2003). Cardiovascular disease among Alaska Natives: a review of the literature. *Int J Circumpolar Health 62*, 343-362.

Statistics Canada; Aboriginal Population Profile, (http://www.statcan.gc.ca: http://www.statcan.gc.ca) [March 1, 2013 accessed].

Stenson, P.D., Ball, E.V., Howells, K., Phillips, A.D., Mort, M., and Cooper, D.N. (2009). The Human Gene Mutation Database: providing a comprehensive central mutation database for molecular diagnostics and personalized genomics. *Hum Genomics 4*, 69-72.

Stoner, L., Stoner, K.R., Young, J.M., and Fryer, S. (2012). Preventing a Cardiovascular Disease Epidemic among Indigenous Populations through Lifestyle Changes. *Int J Prev Med 3*, 230-240.

Sun, P., Dwyer, K.M., Merz, C.N., Sun, W., Johnson, C.A., Shircore, A.M., and Dwyer, J.H. (2000). Blood pressure, LDL cholesterol, and intima-media thickness: a test of the "response to injury" hypothesis of atherosclerosis. *Arterioscler Thromb Vasc Biol 20*, 2005-2010.

Teslovich, T.M., Musunuru, K., Smith, A.V., Edmondson, A.C., Stylianou, I.M., Koseki, M., Pirruccello, J.P., Ripatti, S., Chasman, D.I., Willer, C.J., *et al.* (2010). Biological, clinical and population relevance of 95 loci for blood lipids. *Nature 466*, 707-713.

Ward, H., Mitrou, P.N., Bowman, R., Luben, R., Wareham, N.J., Khaw, K.T., and Bingham, S. (2009). APOE genotype, lipids, and coronary heart disease risk: a prospective population study. *Arch Intern Med 169*, 1424-1429.

Young, T.K., Moffatt, M.E., and O'Neil, J.D. (1993). Cardiovascular diseases in a Canadian Arctic population. *Am J Public Health 83*, 881-887.

Yu, C.H., and Zinman, B. (2007). Type 2 diabetes and impaired glucose tolerance in aboriginal populations: a global perspective. *Diabetes Res Clin Pract 78*, 159-170.

CHAPTER 3

GENETIC DETERMINANTS OF "COGNITIVE IMPAIRMENT, NO DEMENTIA"

The work in this chapter originates from material in the following publication: **Dubé, J.B**., Johansen, C.T., Robinson, J.F., Lindsay, J., Hachinski, V., and Hegele, R.A. (2013). Genetic determinants of "cognitive impairment, no dementia". *J Alzheimers Dis 33*, 831-840.

3.1 INTRODUCTION

Dementia is primarily a disease of the elderly defined by insidious cognitive decline that impairs social and occupational functioning (American Psychiatric Association. Task Force on DSM-IV., 1994; Burns and Iliffe, 2009; Feldman et al., 2008; Geldmacher and Whitehouse, 1996). An early phenotype of cognitive decline that affects ~10-20% of elderly populations (Di Carlo et al., 2007; Graham et al., 1997) is called "cognitive impairment, no dementia" (CIND). CIND is defined broadly by subtle deficiencies in memory or executive functioning that do not fit the definition of dementia but are also abnormal (Chertkow et al., 2008; Tuokko et al., 2001). These findings are commonly associated with increased susceptibility to more severe, later stages of dementia (Tuokko et al., 2003) thus a thorough understanding of the pathogenesis of CIND could lead to improved methods for identifying at-risk patients.

The late stages of dementia are most commonly associated with AD- and vascular dementia (VaD)-related mechanisms of pathogenesis. The prevailing models of AD pathogenesis have implicated perturbations in lipid metabolism, intracellular trafficking and inflammatory pathways that are associated with hallmarks such as β-amyloid (Aβ) plaque deposition and neurofibrillary tangles (NFTs) (Holtzman et al., 2012; Huang and Mucke, 2012). The mechanisms underlying VaD are defined by progressive degeneration or occlusion of the cerebrovasculature that is believed to contribute to a microenvironment of hypoxia and inflammation in the brain (Ballard et al., 2004; Cechetto et al., 2008; Gorelick et al., 2011; Kalaria, 2000; Wolf, 2012). Positive association between AD or cardio-metabolic traits with CIND susceptibility may implicate the effects of established late-stage degenerative mechanisms at an earlier stage of cognitive decline.

Genetic variation has been associated with both AD (Hollingworth et al., 2011; Naj et al., 2011) and the cardiovascular traits associated with vascular disease (Hegele, 2009; Teslovich et al., 2010), however no studies have tested for association between these variants and CIND susceptibility. Genome-wide association studies (GWAS) (Hirschhorn and Daly, 2005) are a method of testing for association between genetic variation and a heritable trait that has successfully identified novel genes involved in vascular health such as blood lipid traits (Hegele, 2009; Teslovich et al., 2010) or disease phenotypes such as Alzheimer-related dementia (Hollingworth et al., 2011; Naj et al., 2011). A GWAS-based approach applied to CIND could reveal whether genes implicated in VaD and AD susceptibility are also associated with CIND. Here, we conduct a GWAS

to determine whether genetic variation previously associated with cardio-metabolic traits or AD was associated with CIND susceptibility. We genotyped ~200,000 genetic variants associated with multiple metabolic traits using the Cardio-MetaboChip (MetaboChip), and 12 genetic variants strongly associated with AD susceptibility (including the APOE isoform). We provide the first comprehensive genetic evaluation of CIND and demonstrate a novel locus potentially increasing CIND susceptibility.

3.2 MATERIALS AND METHODS

3.2.1 Study cohort

This study was approved by University of Western Ontario Institutional Review Board (Review number 07920E). The Canadian Study of Health and Aging (CSHA) was a longitudinal population-based cohort study designed to document the prevalence of dementia and related variables in elderly Canadian communities from 1991 to 2001 (n=10,263) (1994). CSHA participants were aged ≥65 years at the start of the study. CIND patients (n=528) were selected based on clinical diagnoses of CIND based on Diagnostic and Statistical Manual of Mental Disorders (DSM-IV) criteria (American Psychiatric Association. Task Force on DSM-IV., 1994) excluding patients with presumed alcohol or drug use, mental retardation or other psychiatric illness. CIND diagnoses were adjudicated by a panel of neuropsychologists and physicians during the course of the CSHA, whereas control subjects (n=494) were selected from cognitively normal CSHA participants. Modified Mini-Mental State Exam (3MS) (Teng and Chui,

1987) scores obtained during the CSHA confirmed CIND and control status in patients selected for this study.

3.2.2 Study design

We used a two-stage GWAS design comprising discovery and replication phases. The discovery phase involved genotyping a cohort of CIND patients (n=274) and controls (n=301) on the MetaboChip – a custom Illumina genotyping array (Illumina Inc.; San Diego, CA) populated with ~200,000 single nucleotide polymorphisms (SNPs) identified from GWAS of metabolic traits and cardiovascular phenotypes (http://www.sph.umich.edu/csg/kang/MetaboChip/). MetaboChip genotyping was performed at the Broad Institute (http://www.broadinstitute.org/). Variants included in our analysis had call-rates >90%, minor allele frequency (MAF) >1% and Hardy Weinberg P>10-4. The replication phase involved genotyping an independent cohort of CIND patients (n=210) and controls (n=158) for the 13 most significant loci identified by the MetaboChip. These variants were genotyped using either TaqMan SNP genotyping assays (Applied Biosystems; Foster City, CA) or direct Sanger sequencing (**Table 3.1**). A cohort of our study (339 cases, 304 controls) was also genotyped for the top 11 AD-associated variants using TaqMan genotyping assays (**Table 3.1**). The same quality control filters applied in the discovery phase were applied to all genotyped variants.

3.2.3 Statistical analyses

Study cohort demographics were evaluated using chi-square ($\chi 2$) tests for dichotomous variables and t-tests for continuous variables using SAS v9.2 (Cary, NC); statistical

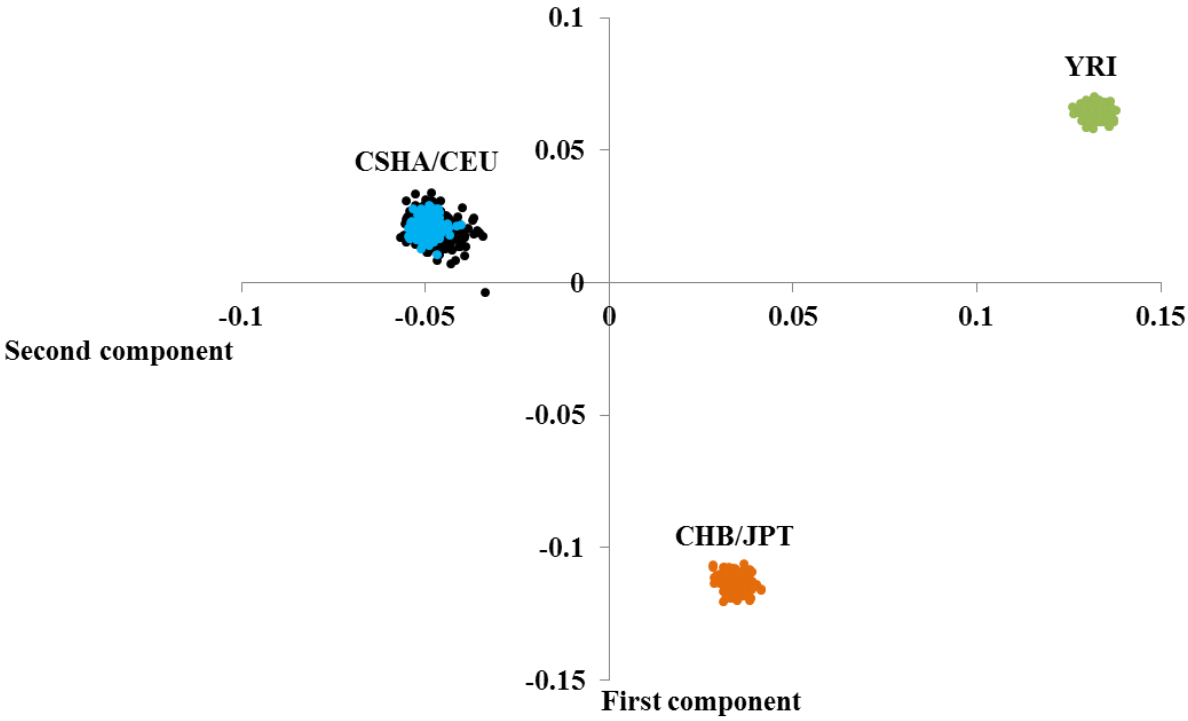
Table 3.1 Custom variant genotyping assays and primer designs.

MetaboChip-related SNP genotyping assays				
Gene	**SNP**	**TaqMan**	**Primer sequence**	**Conditions**
FLJ22536	rs16901621	N	5'-TCCTTTCCAGGGTGCAAGTC-3' 5'-ATCTTCATTCAGCCCCAGAC-3'	58^{o}C annealing
IRX1	rs13186537	N	5'-ATTGAGGATTCATTTGTGGC-3' 5'-AGCCCCTGTTTTTTACCTGTC-3'	56^{o}C annealing
ME1	rs1145909	N	5'-CCATCCCACATTTATGCAG-3' 5'-GCTTTTGGCAACCACTTTCTAG-3'	58^{o}C annealing
EHD4	rs1704405	Y	5'-TCAGACTTCACAAAGTGGGAATTTGA-3' 5'-TTGCCACTGCCCTTTGTCT-3'	NA
Alzheimer-related SNP genotyping assays				
Gene	**SNP**	**TaqMan**	**Primer sequence**	**Conditions**
CD2AP	rs9349407	Y	5'-AATGTAGTTAGCTTTAGTGTATGGTGTTTATAAAATCT-3' 5'-CAGTGAGTGGTGAGCAAATGTG-3'	NA
CD33	rs3865444	Y	5'-GAGTCGCAGCCTCACCTA-3' 5'-CTCACACGGACCCTATAGAATCCTA-3'	NA

SNPs not listed here were genotyped using pre-designed TaqMan SNP genotyping assays. The manufacturer's suggested assay conditions were used. NA, not applicable.

significance was defined as $P<0.05$. Logistic regression was used to test for association between genetic variants and CIND status using PLINK (http://pngu.mgh.harvard.edu/purcell/plink/) (Purcell et al., 2007). Statistical significance in the discovery phase was defined as a Bonferroni-corrected $P<4.0x10^{-7}$. Our logistic regression model was adjusted for potentially confounding variables including age, sex, years of education, APOE ε4 carrier status, history of stroke or possible stroke and two principal components of ancestry. The significance threshold for the replication phase was set at $P<3.8x10^{-3}$. Significance values reported for the combined cohort were used only to show the change in significance with increasing sample size. It was not possible to adjust for ancestry in the replication cohort; however it was unlikely to play a role. We compared CSHA participant ancestry with HapMap populations of known ancestry (2003) using identity-by-state and multidimensional scaling based on >50,000 variants to confirm ancestry reported during the CSHA (**Figure 3.1**). Manhattan plots and quantile-quantile (Q-Q) plots were used to visualize results using WGAViewer (Ge et al., 2008). AD genetic risk scores (AD-GRS) were constructed using 11 AD-associated variants identified from the largest AD GWAS meta-analyses (Hollingworth et al., 2011; Naj et al., 2011). For each variant, 1 allele has been associated with AD risk (Odds ratio, OR, >1.0) therefore in each participant, we counted the number of copies of risk alleles which comprised a composite risk score. Risk alleles for the 11 AD-associated non-APOE variants were identified based on data from the same meta-analyses (Bertram et al.). The difference between mean AD-GRSs in cases and controls was tested using an independent samples t-test defining statistical significance as $P<0.05$.

Figure 3.1 Principal components analysis with Canadian Study of Health and Aging (CSHA)- and HapMap-derived populations. The first two principal components were plotted with HapMap populations of known ancestry to confirm the reported ancestry of CSHA participants. CEU (blue, n = 165), Caucasians; CSHA (black, n = 575), CSHA discovery phase cohort; CHB/JPT (orange, n = 250), Chinese/Japanese; YRI (green, n = 203), African.


0.1
0.05
0
-0.05
-0.1
-0.15
-0.1
-0.05
0
0.05
0.1
0.15
YRI
CSHA/CEU
CHB/JPT
Second component
First component

3.2.4 Power calculations

For the discovery phase of our GWAS, we calculated a >80% probability of rejecting the null hypothesis of no association for a common variant with strong biological effect (OR=2.50, MAF=0.20) using a genome-wide level of Type I error probability. Based on sample size in the replication phase and results from the GWAS discovery phase, we estimated a <5% probability of rejecting the null hypothesis of no association for variants of expected effect size and frequency (OR=1.50, MAF=0.30) using the Bonferroni-corrected P-value of 3.8×10^{-3}. When testing for association between AD-associated variants and CIND, we calculated ~20% probability of rejecting the null hypothesis of no association for a common variant with modest biological effect (OR=1.20, MAF=0.40) using the standard level of Type I error probability. Power calculations were performed using Power and Sample Size Calculation software (Dupont and Plummer, 1998).

3.3 RESULTS

3.3.1 Study subjects

Demographic data for CSHA participants selected for this study are shown in **Table 3.2**. Cases and controls differed most significantly in 3MS score ($P=6.8 \times 10^{-38}$) and incident stroke ($P=6.9 \times 10^{-4}$). Cases also had non-significantly higher frequencies of the remaining indices of cardiovascular disease (CVD) compared to controls.

Table 3.2 Study cohort demographics for CSHA controls and cases.

	Combined		Discovery phase		Replication phase		
	Controls	**CIND**	**Controls**	**CIND**	**Controls**	**CIND**	**P-value**
n	459	484	301	274	158	210	
Male (%)	45	45	46	46	45	44	NS
Age (±SD)	75.6±6.6	76.4±6.5	76.2 ± 6.6	77.5±6.6	74.4±6.5	75.1±6.2	NS
Education (years, ±SD)	10.1±4.0	9.7±5.8	10.2±5.6	9.3±5.6	9.9±3.8	10.1±3.9	NS
3MS (±SD)	88.0±7.5	80.2±9.9	88.5±9.6	81.8±9.6	88.5±6.3	83.1±9.0	6.8×10^{-38}
Stroke (%)	6.5	13.0	8.3	18.2	3.2	6.7	6.9×10^{-4}
Diabetes mellitus (%)	12.6	16.9	14.0	16.1	10.1	18.1	0.07
Hypertension (%)	50.8	55.2	49.5	52.9	53.2	58.1	NS
Circulation problems (%)	32.7	35.7	37.5	40.5	34.0	36.7	NS

P-values are based on the combined cohort. 3MS, modified mini-mental state examination; NS, not significant (P-value>0.05); SD, standard deviation.

3.3.2 GWAS of CIND

First, we tested for association between genetic variation in cardio-metabolic genes and CIND status using the MetaboChip (**Figure 3.2**). The most significant CIND-associated variant from the discovery phase was rs16901621 in *FLJ22536* (OR=2.67; $P=3.2x10^{-7}$) which modestly surpassed the Bonferroni-corrected threshold of significance ($P<4.0x10^{-7}$). Given that no variants achieved stringent Bonferroni-corrected significance thresholds, with exception to the FLJ22536 variant, we selected the 13 most significant variants ($P<1.0x10^{-3}$) for replication in an independent cohort (**Table 3.3**). We identified a locus near *ZNF608/GRAMD3* (rs1439568) approaching statistical significance for replication (OR=0.66; $P=6.0x10^{-3}$), however we were unable to replicate the initial finding in *FLJ22536* (OR=0.82; P=0.27) (**Table 3.3**). Although we observed similar effect sizes between discovery and replication phase variants, no variants absolutely surpassed a Bonferroni-adjusted threshold of significance ($P<4.0x10^{-7}$). The rs1439568 polymorphism lies within a ~4kb haplotype block void of well-annotated genes, located ~500kb downstream of *ZNF608* and ~115kb upstream of *GRAMD3* (**Figure 3.3**). This locus may represent a putative genetic determinant of CIND susceptibility.

3.3.3 AD-associated variation in CIND

Next, we sought to assess whether AD-related genetic variation was associated with CIND. We genotyped 11 AD-associated variants reported in recent GWAS meta-analyses of AD-related dementia (Hollingworth et al., 2011; Naj et al., 2011). No significant associations were identified between any of the previously defined risk alleles, although

Figure 3.2 Manhattan plot showing results from the MetaboChip genome-wide association study in the discovery phase. Manhattan plots help visualize the loci strongly associated with disease susceptibility by plotting all SNPs together based on association P-value and physical position. Each point represents a P-value of a test for association between a SNP and CIND status. The significance of association is plotted on the y-axis with increasing significance ranking higher on the y-axis. The genomic position of the SNP corresponding to each test for association is plotted on the x-axis. Polymorphisms with P-values $<10^{-3}$ are shown in red.

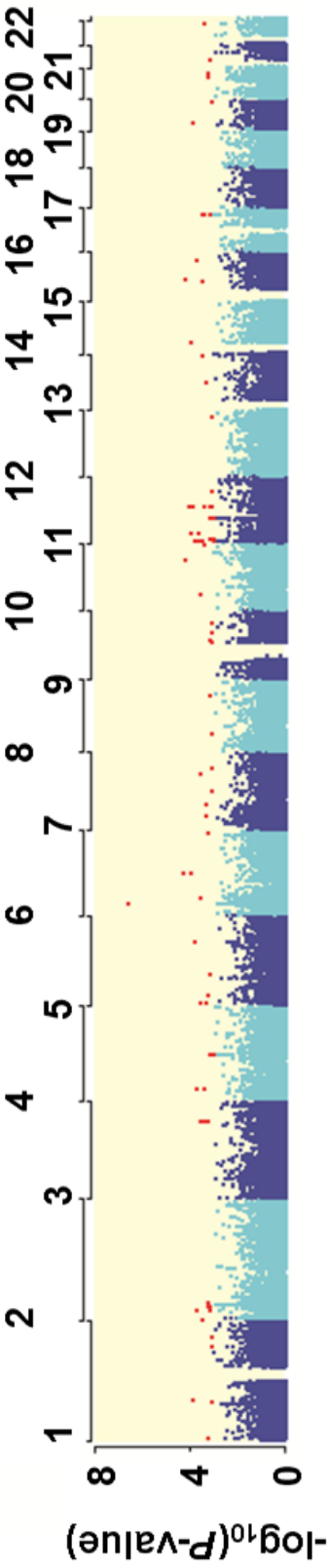
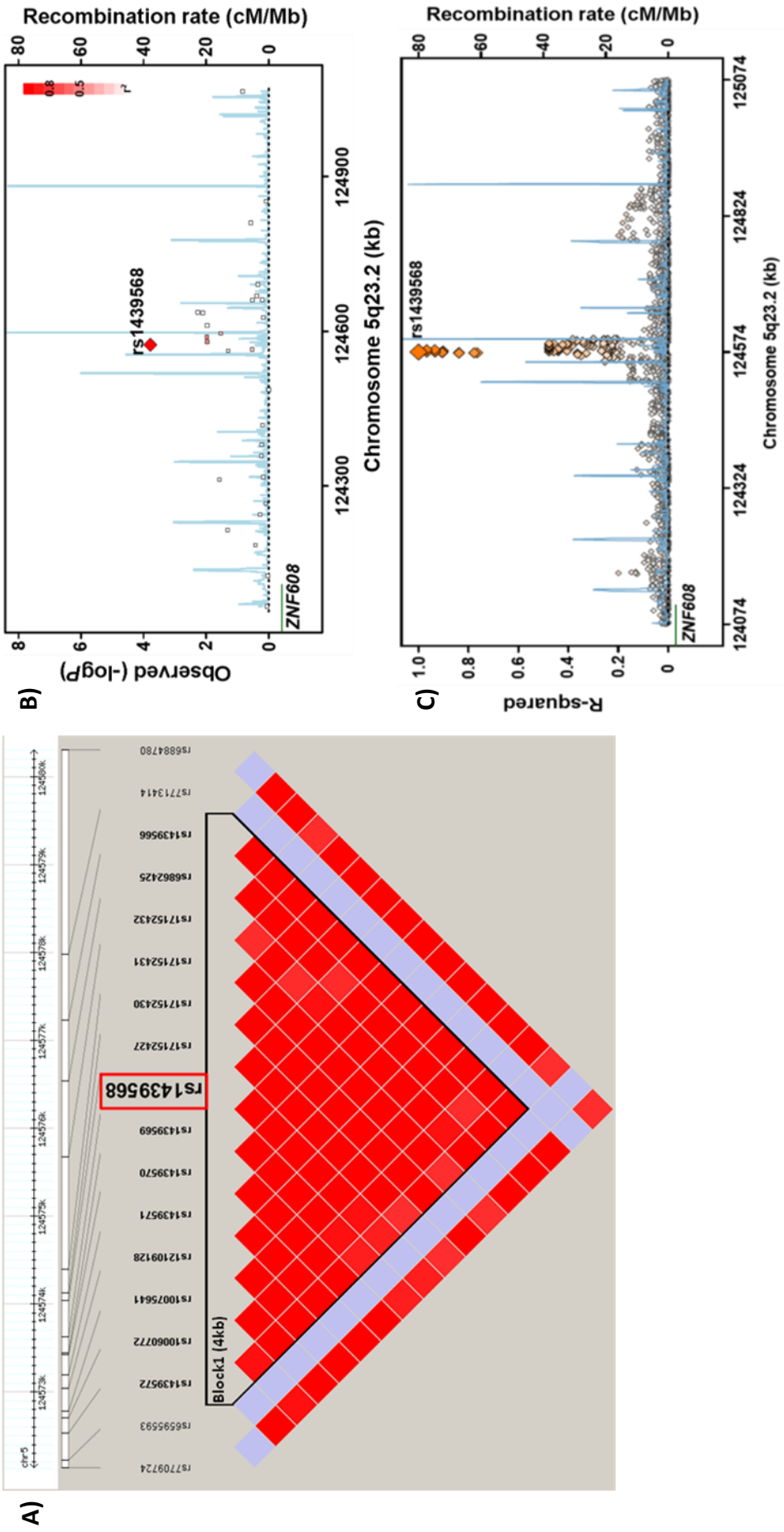

-log10(P-value)
0
4
8
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22

Table 3.3 Results from association tests between top MetaboChip variants and CIND status in discovery and replication phases.

CHR	SNP	Nearest gene (trait)	Allele (min/maj)	MAF	Discovery		Replication		Combined	
					OR (95% CI)	P	OR (95% CI)	P	OR (95% CI)	P
5	rs1439568	*ZNF608/GRAMD3* (LDL)	G / A	0.44	0.62 (0.49-0.80)	$1.60x10^{-4}$	0.66 (0.49-0.89)	0.0060	0.66 (0.54-0.79)	$8.36x10^{-6}$
15	rs1704405	*EHD4* (QT)	G / A	0.22	1.91 (1.40-2.60)	$4.48x10^{-5}$	1.24 (0.88-1.74)	0.22	1.54 (1.23-1.93)	$1.66x10^{-4}$
10	rs556474	*CRTAC1* (DBP)	A / G	0.42	1.65 (1.29-2.12)	8.4610^{-5}	1.14 (0.85-1.52)	0.40	1.43 (1.18-1.72)	$2.02x10^{-4}$
11	rs11023937	*KCNQ1* (QT)	G / C	0.33	0.59 (0.45-0.77)	$1.40x10^{-4}$	0.81 (0.59-1.10)	0.18	0.69 (0.57-0.85)	$3.23x10^{-4}$
1	rs7518019	*LOC553139* (2HR GLU)	G / A	0.05	0.28 (0.14-0.53)	$1.30x10^{-4}$	0.80 (0.34-1.84)	0.60	0.41 (0.25-0.66)	$3.31x10^{-4}$
5	rs13186537	*IRX1* (MI/CAD)	A / C	0.07	2.68 (1.62-4.45)	$1.30x10^{-4}$	1.26 (0.71-2.23)	0.43	1.99 (1.36-2.92)	$3.95x10^{-4}$
6	rs1145909	*ME1* (SBP)	A / C	0.16	0.48 (0.33-0.68)	$4.62x10^{-5}$	0.87 (0.59-1.30)	0.50	0.64 (0.50-0.83)	$6.39x10^{-4}$
11	rs2615016	*SOX6* (LDL)	C / G	0.48	1.66 (1.29-2.13)	$6.77x10^{-5}$	1.08 (0.81-1.46)	0.59	1.35 (1.12-1.62)	$1.71x10^{-3}$
6	rs16901621	*FLJ22536* (DBP)	G / A	0.16	2.67 (1.83-3.90)	$3.22x10^{-7}$	0.82 (0.58-1.17)	0.27	1.49 (1.16-1.92)	$1.93x10^{-3}$
11	rs10898893	*FCHSD2* (NR)	G / A	0.18	1.90 (1.39-2.59)	$5.81x10^{-5}$	0.92 (0.64-1.33)	0.67	1.44 (1.14-1.82)	$2.18x10^{-3}$
7	rs10275038	*DOCK4* (FAST GLU)	A / G	0.20	0.55 (0.40-0.75)	$1.8x10^{-4}$	1.05 (0.72-1.53)	0.79	0.72 (0.57-0.91)	$6.43x10^{-3}$
2	rs7574887	*GEN1* (MI/CAD)	G / A	0.09	2.28 (1.48-3.52)	$1.90x10^{-4}$	0.88 (0.55-1.41)	0.59	1.51 (1.11-2.07)	$9.57x10^{-3}$
6	rs13218698	*DOPEY1* (2HR GLU)	C / A	0.16	0.48 (0.34-0.69)	$6.66x10^{-5}$	1.12 (0.77-1.62)	0.56	0.74 (0.58-0.95)	0.019

ORs and P-values are based on the minor allele. Risk alleles (OR>1.0) are underlined. 2HR GLU, two-hour blood glucose level; CHR, chromosome; DBP, diastolic blood pressure; FAST GLU, fasting blood glucose level; LDL, low-density lipoprotein cholesterol level; MAF, minor allele frequency; maj, major (more frequent) allele; min, minor (less frequent) allele; MI / CAD, myocardial infarction/coronary artery disease; NR, not reported; QT, QT interval; SBP, systolic blood pressure; SNP, single nucleotide polymorphism.

Figure 3.3 Regional genetic variation in the vicinity of rs1439568. A) Haplotype block surrounding rs1439568 based on data from the HapMap CEU population. Red boxes represent a high degree of linkage disequilibrium (LD) between two markers whereas grey boxes suggest weak linkage disequilibrium. The relative position of rs1439568 is outlined in red. By investigating LD between a variant of interest and additional nearby variants, it is possible to better define the boundaries of a potential disease susceptibility locus and whether regulatory elements fall within this locus such as a promoter region or transcription factor binding sites. B) LD between MetaboChip genotyped SNPs and rs1439568. Discovery phase-calculated p-values (left y-axis) for SNP association with CIND determine the height of each point. The degree of LD between a SNP and rs1439568 is proportional to the intensity of red colouration. Blue peaks identify sites of recombination (right y-axis). When visualized together, this plot helps identify whether the disease-associated SNP is in LD with other local SNPs and whether the local SNPs were also associated with disease status. C) Regional LD between rs1439568 and neighboring SNPs within 500 kilobases of rs1439568 based on the HapMap CEU dataset. Each point represents a SNP and its height on the left y-axis indicates the strength of linkage disequilibrium between a particular SNP and rs1439568. Blue peaks correspond to sites of recombination (right y-axis). This plot provides an enhanced visualization of local SNPs in LD with rs1439568 as this locus was more densely genotyped by the HapMap consortium. Data from panels A) and C) were generated using the hg18 build.


A)
rs1439568
Block1 (4kb)
B)
Recombination rate (cM/Mb)
rs1439568
ZNF608
Observed (-logP)
Chromosome 5q23.2 (kb)
C)
Recombination rate (cM/Mb)
rs1439568
ZNF608
R-squared
Chromosome 5q23.2 (kb)

the effect sizes for some variants, including *CR1*, *ABCA7*, and *PICALM*, were consistent with previously published studies (**Table 3.4**). We also constructed a composite AD-GRS using risk-increasing alleles from the AD-associated variants to assess an overall difference in the accumulation of multiple AD-associated genetic variants between CIND patients and controls (**Figure 3.4**). The frequencies of AD-GRSs between CIND patients and controls were not significantly different when we compared mean scores, which totaled ~9 risk alleles in CIND patients and controls (P=0.71). This suggests that putative AD-associated variants are not involved in CIND predisposition.

3.3.4 APOE status in CIND

Finally, we evaluated the frequency of the APOE E4 allele in CIND patients versus cognitively healthy controls. The frequency of the APOE E4 allele was elevated in CIND patients versus controls. Each copy of the APOE E4 allele increased CIND susceptibility compared to the APOE E3 allele (OR=1.35; P=0.044) (**Table 3.5**). The APOE E2 allele frequency was non-significantly higher in cases versus controls (OR=1.14; P=0.44).

3.4 DISCUSSION

The principal finding of our study is that non-APOE genetic variation associated with cardio-metabolic traits or AD is not associated with CIND susceptibility. We were not able to identify any loci associated with CIND through a GWAS of >200,000 cardio-metabolic-associated variants, nor were we able to replicate the top 11 non-APOE variants associated with AD susceptibility either individually or as part of an AD-GRS.

Table 3.4 Results from association tests between Alzheimer disease-associated variants and CIND status.

CHR	SNP	Nearest gene	Allele (min/maj)	MAF	Reported OR (95%CI)	OR (95% CI)	P
1	rs3818361	*CR1*	A / G	0.18	1.16 (1.11-1.22)	1.27 (0.94 - 1.71)	0.12
8	rs11136000	*CLU*	T / C	0.41	1.12 (1.10-1.16)	0.90 (0.71 - 1.14)	0.39
19	rs3764650	*ABCA7*	G / T	0.09	1.23 (1.18-1.28)	1.14 (0.80 - 1.64)	0.47
11	rs3851179	*PICALM*	T / C	0.36	1.14 (1.09-1.16)	1.08 (0.86 - 1.37)	0.47
11	rs670139	*MS4A4E*	T / G	0.38	1.08 (1.05-1.10)	0.92 (0.74 - 1.15)	0.47
11	rs610932	*MS4A6A*	T / G	0.43	1.11 (1.08-1.14)	1.05 (0.76 - 1.19)	0.67
6	rs9349407	*CD2AP*	G / C	0.28	1.12 (1.07-1.17)	1.04 (0.82 - 1.32)	0.75
19	rs3865444	*CD33*	A / C	0.32	1.12 (1.09-1.16)	0.98 (0.78 - 1.23)	0.88
19	rs597668	*EXOC3L2*	C /T	0.14	1.17 (1.12-1.23)	1.01 (0.74 - 1.38)	0.95
7	rs11767557	*EPHA1*	C / T	0.19	1.12 (1.04-1.20)	1.00 (0.75 - 1.32)	0.99
2	rs744373	*BIN1*	G / A	0.29	1.17 (1.13-1.20)	1.00 (0.80 - 1.26)	0.99

Underlined alleles represent the previously reported risk alleles from genome-wide association study meta-analyses. Reported ORs represent meta-analyses from the AlzGene database (Bertram et al., 2013). Abbreviations as in Table 3.3.

Figure 3.4 Frequency distribution of Alzheimer disease (AD) genetic risk scores in cognitive impairment no dementia (CIND) patients and controls. Risk alleles from 11 non-APOE AD-associated variants were added in each study participant. For each variant, a participant can have 0, 1, or 2 copies of the risk allele thus scores could range from 0–22. Risk scores were calculated in CIND patients and controls. The difference between the mean risk scores for CIND patients and controls was not different (P=0.71).

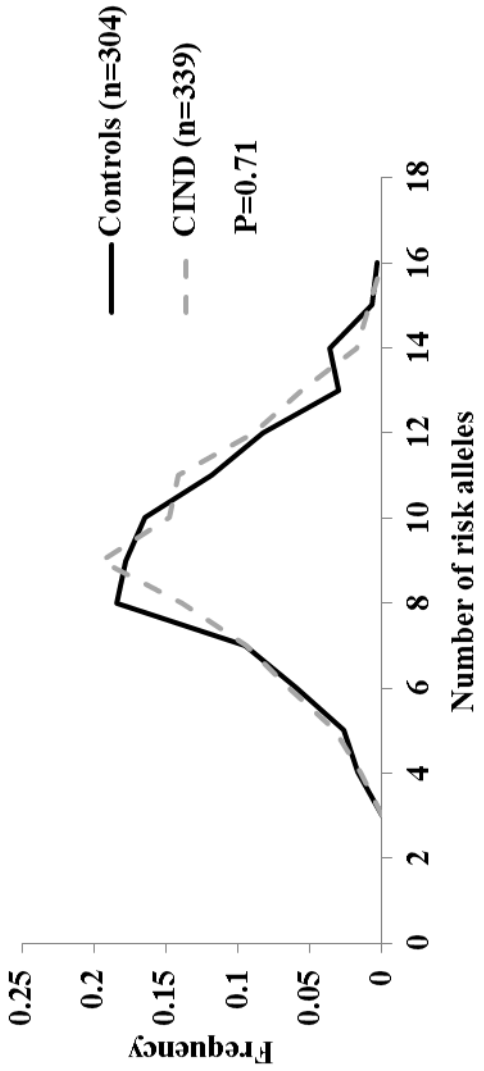

Controls (n=304)
CIND (n=339)
P=0.71
Frequency
0.25
0.2
0.15
0.1
0.05
0
0
2
4
6
8
10
12
14
16
18
Number of risk alleles

Table 3.5 *APOE* allele frequencies in CIND cases and controls.

***APOE* allele**	**Controls[1] (%)**	**CIND[2] (%)**	**OR (95% CI)**	***P*-value**
E2	85 (8.6)	99 (9.4)	1.14 (0.83-1.56)	0.44
E4	92 (9.3)	127 (12.0)	1.35 (1.00-1.81)	0.044
E3	807 (82.0)	828 (78.6)	NA	NA

[1]n=492; [2]n=527. Odds ratios (ORs) were calculated based on *APOE* allele frequency relative to *APOE* E3 frequency. *APOE*, apolipoprotein E; CI, confidence interval; CIND, "cognitive impairment no dementia"; NA, not applicable.

Despite these negative findings, we confirmed that the APOE E4 allele increases CIND susceptibility. These contributions provide the most comprehensive genetic analysis of CIND susceptibility conducted to date.

Our results provide some insight into the genetic basis of CIND. Our study explicitly shows that the APOE E4 allele is associated with CIND susceptibility. Previous studies have primarily evaluated CIND in the context of progression to late-stage dementia phenotypes such as AD. For instance, one study of 68 CIND patients conducted independently by the CSHA showed that APOE E4 carriers were 2.7-times more likely to progress to AD (Hsiung et al., 2004). Other studies evaluating the APOE E4 allele in the context of pre-dementia have used a phenotype called mild cognitive impairment (MCI) that is known to progress to AD (Albert et al., 2011; Feldman and Jacova, 2005). Meta-analysis of 35 studies including 6095 MCI patients and 1236 AD patients reported an association between the E4 allele in MCI to AD progression ($P<0.001$, OR=2.29, 95% CI=1.88-2.80) (Elias-Sonnenschein et al., 2011). Our study is distinct in that we have identified a similar but relatively modest association between the E4 allele and CIND susceptibility in 527 CIND patients and 492 cognitively healthy controls ($P=0.044$, OR=1.35, 95% CI = 1.00-1.81). Our data suggest that APOE E4 increases susceptibility to CIND regardless of progression to late stage dementia, although with an effect size approximately half of what is normally reported with disease progression. Such differences in effect size may represent differences in sample size between studies, or rather it may represent the underlying phenotypic heterogeneity within the clinical definition of CIND, as discussed below. More careful evaluation of APOE is clearly

needed to determine how the APOE E4 allele influences susceptibility to different forms of cognitive decline.

We also show that VaD-related genetic variation is not associated with CIND. Since VaD represents the second most common form of late-onset dementia, we used a custom SNP array called the MetaboChip to investigate whether genetic loci associated with vascular health including coronary artery disease, type 2 diabetes, and plasma lipid concentrations were also associated with CIND. We were surprised to find no significant association between genetic variation at cardio-metabolic loci and CIND since vascular disease has an established correlation with cognitive decline (Gorelick et al., 2011). Based on previous genetic studies in VaD, however, our findings may not be entirely unexpected. A twin study assessing VaD showed little pair-wise concordance for VaD suggesting a diminished role for genetics and a greater contribution of environmental factors in VaD risk (Bergem et al., 1997). Furthermore, the only VaD GWAS has failed to identify any loci surpassing a genome-wide threshold of significance ($P<10^{-8}$) thus it is possible that the variants associated with cardio-metabolic traits do not significantly affect cognitive health or CIND susceptibility (Schrijvers et al., 2011).

Interestingly, we identified one variant near the *ZNF608/GRAMD3* locus that was nominally associated with CIND (P=0.0060, OR=0.66, 95% CI=0.49-0.89), although it did not surpass stringent thresholds for association. *ZNF608* was previously associated with LDL cholesterol, which facilitated mapping of this locus on the MetaboChip. GWAS have reported associations of common variants at *ZNF608* with neuroblastoma

and obesity (Speliotes et al., 2010). Differential *ZNF608* expression has also been observed in the prefrontal cortex during fetal and infant development (Colantuoni et al., 2011). *ZNF608* encodes a modestly characterized zinc-finger motif protein where functional studies of the Drosophila melanogaster ZNF608 homologue, scribbler (sbb), suggest a role for ZNF608 in transcriptional repression and starvation resistance (Harbison et al., 2004). Additionally, ZNF608 contains ubiquitylation sites which suggest a regulation of ZNF608 stability via post-transcriptional modification (PTM) (Wagner et al., 2011). Conversely, our current understanding of the biological function of GRAMD3 is limited, thus the role of common variation at *ZNF608/GRAMD3* in CIND susceptibility is speculative at this point. Further association studies must be performed to confirm this potential CIND susceptibility locus while sequencing of the *ZNF608/GRAMD3* locus for rare variant analysis may reveal insight into biological relevance by the identification of functional variants.

Lastly, we were unable to extend associations previously reported between non-APOE genetic variants associated with AD and the CIND phenotype. As a previous study from the CSHA showed that CIND patients were 5 times more likely to develop AD than cognitively normal participants (OR=5.0, 95% CI=3.4-7.3) (Tuokko et al., 2003), we sought to test whether the top non-APOE AD-associated variants were associated with CIND status. We found no significant association between established AD-associated variants and CIND which was also confirmed in our AD-GRS analysis suggesting that AD-related mechanisms of cognitive decline may not be a strong determinant of CIND. The association between the APOE E4 allele and CIND proposes one commonality

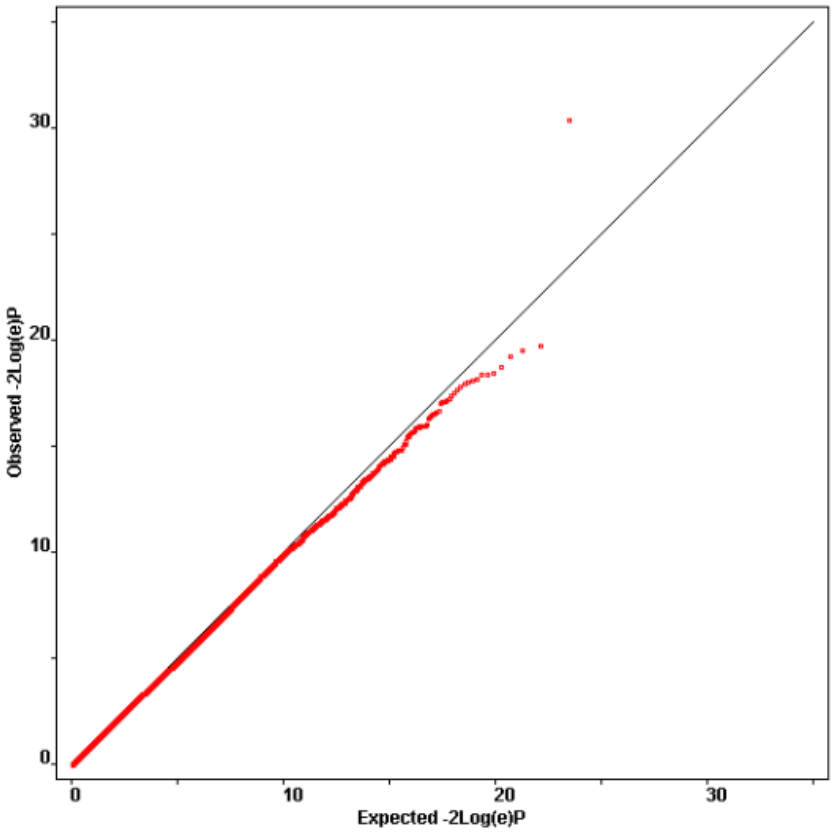
between CIND and AD, however, the effect size associated with the E4 allele in our CIND cohort (OR=1.35, 95% CI=1.00-1.81) pales in comparison to the effect size associated with the E4 allele in an AD meta-analysis of 37 studies (OR=3.68, 95% CI=3.30-4.11) (Bertram et al.). Although we did not find strong evidence supporting a similar genetic architecture between CIND and AD, an APOE-related pathway may be an important determinant of susceptibility to cognitive decline in CIND.

A significant study limitation pertains to the phenotypic heterogeneity inherent in CIND and pre-dementia. A spectrum of psychiatric diseases is associated with dementia which extends to the onset of cognitive decline where a myriad of disease mechanisms may initiate cognitive decline (Tarawneh and Holtzman, 2012; Tuokko et al., 2001). Post-mortem studies of dementia patients commonly report mixed cerebrovascular and AD-associated pathologies in patient brains, which supports a synergistic role for multiple disease mechanisms in cognitive decline (Schneider et al., 2007). The spectrum of disease severity among CIND patients further complicates the process of selecting an ideally homogenous CIND cohort, since CIND patients have the potential to steadily decline, remain stable or even improve cognitively (Tuokko et al., 2001). Within our study cohort, we lacked the end-stage phenotyping required to diagnose a sub-type of dementia such as VaD or AD, however, our study focused on the genetics of CIND irrespective of dementia subtypes thus end-stage diagnosis was not a hindrance in our study. In contrast, other studies that seek to characterize the transition from pre-dementia to defined end-stage diseases will rely on the availability of prospective data and end-stage phenotyping. The development of the MCI phenotype mirrors this logic as MCI

was conceptualized as a pre-dementia stage specific to AD patients. Thus AD diagnosis must be considered when studying MCI (Tarawneh and Holtzman, 2012).

Our study also had limited statistical power to detect significant associations with small effect sizes. While our study of AD GWAS variants in CIND was negative, it is important to note that the established non-APOE variants typically have small ORs <1.20. Meta-analysis involving thousands of AD patients and controls was required to achieve adequate statistical power to identify small but significant effect sizes; our study was markedly underpowered to detect such effect sizes. In order to adapt our study to a limited sample size, we implemented a two-stage GWAS approach in which we only tested for association between cardio-metabolic loci within a population showing normality according to our Q-Q plot (**Figure 3.5**). These approaches improved our ability to detect a significant association by lowering the stringent statistical requirements typical of true GWAS, however, the issue of limited statistical power was not ameliorated. Based on the relatively larger effect size associated with the APOE E4 allele in AD, our study was adequately powered to detect a significant association assuming a similar effect size in CIND patients as observed in AD patients. Our AD GRS approach also lowered statistical stringency as a $P<0.05$ specifies significance rather than a Bonferroni-adjusted P-value cutoff applied to the individual tests for association (approx. $P<0.0045$). In order to address the issue of power, larger CIND cohorts must be organized, however ours is the largest genetic study of CIND patients conducted to date. A sample size of ~3000 CIND patients and 3000 healthy controls would be required to detect a significant association at the AD GWAS-identified *BIN1* locus assuming a

Figure 3.5 Quantile-quantile plot showing expected and observed p-values from the MetaboChip discovery phase. QQ plots visualize the difference between the observed P-values for each SNP derived from the GWAS discovery phase compared with expected P-values derived from a theoretical χ^2 distribution. The solid diagonal line represents the null hypothesis of no difference between observed and expected P-values. Deviation from the null hypothesis highlights the presence of inflation or deflation in observed P-values which can be caused by the presence of unadjusted population stratification. The distribution of p-values suggests that there was no artificial inflation of test statistics ($\lambda GC = 0.99$).


Observed -2Log(e)P
Expected -2Log(e)P
0
10
20
30
0
10
20
30

consistent effect size (OR=1.17). In comparison, a sample size of only 400 CIND patients and 400 cognitively normal controls is required to replicate the putative association at the *ZNF608/GRAMD3* locus. Thus, it is clear that follow-up studies in CIND and pre-dementia need to be on a similar scale as the scale of published AD GWAS in order to detect associations of small effect. Finally, some limitations can also be attributed to our use of the MetaboChip platform. Since only cardio-metabolic loci were included in our GWAS, it is possible that we overlooked other loci that may contain genetic variation associated with CIND. Imputation of additional variants or the application of a truly genome-wide genotyping platform in a CIND cohort may provide a more comprehensive investigation of genetic variation.

In summary, we have shown that genetic variation in cardio-metabolic and AD-associated loci are not associated with CIND. We identified a potential association between the *ZNF608/GRAMD3* locus and CIND status, however, additional studies are required to validate this association as well as subsequent studies of possible functional effects. Clinical implications arising from this study are hypothesis generating at this stage. However, this novel approach to characterizing cognitive decline may help to implicate a greater role for genetic determinants of cardiovascular traits when applied to larger cohorts. While the synergy between cardiovascular and cognitive health is strong, evidence in this study supporting a genetic link was weak. Identification of genes contributing to cognitive decline via similar approaches as shown here will help us understand the pathways and mechanisms affecting CIND susceptibility which will ultimately guide future therapeutic strategies.

3.5 REFERENCES

Albert, M.S., DeKosky, S.T., Dickson, D., Dubois, B., Feldman, H.H., Fox, N.C., Gamst, A., Holtzman, D.M., Jagust, W.J., Petersen, R.C., *et al.* (2011). The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement 7*, 270-279.

American Psychiatric Association. Task Force on DSM-IV. (1994). Diagnostic and statistical manual of mental disorders : DSM-IV, 4th edn (Washington, D.C., American Psychiatric Association).

Ballard, C.G., Morris, C.M., Rao, H., O'Brien, J.T., Barber, R., Stephens, S., Rowan, E., Gibson, A., Kalaria, R.N., and Kenny, R.A. (2004). APOE epsilon4 and cognitive decline in older stroke patients with early cognitive impairment. *Neurology 63*, 1399-1402.

Bergem, A.L., Engedal, K., and Kringlen, E. (1997). The role of heredity in late-onset Alzheimer disease and vascular dementia. A twin study. *Arch Gen Psychiatry 54*, 264-270.

Bertram, L., McQueen, M., Mullin, K., Blacker, D., and Tanzi, R.E. (2013). The AlzGene Database (Alzheimer Research Forum).

Burns, A., and Iliffe, S. (2009). Dementia. *BMJ 338*, b75.

Cechetto, D.F., Hachinski, V., and Whitehead, S.N. (2008). Vascular risk factors and Alzheimer's disease. *Expert Rev Neurother 8*, 743-750.

Chertkow, H., Massoud, F., Nasreddine, Z., Belleville, S., Joanette, Y., Bocti, C., Drolet, V., Kirk, J., Freedman, M., and Bergman, H. (2008). Diagnosis and treatment of dementia: 3. Mild cognitive impairment and cognitive impairment without dementia. *CMAJ 178*, 1273-1285.

Colantuoni, C., Lipska, B.K., Ye, T., Hyde, T.M., Tao, R., Leek, J.T., Colantuoni, E.A., Elkahloun, A.G., Herman, M.M., Weinberger, D.R., *et al.* (2011). Temporal dynamics and genetic control of transcription in the human prefrontal cortex. *Nature 478*, 519-523.

Di Carlo, A., Lamassa, M., Baldereschi, M., Inzitari, M., Scafato, E., Farchi, G., and Inzitari, D. (2007). CIND and MCI in the Italian elderly: frequency, vascular risk factors, progression to dementia. *Neurology 68*, 1909-1916.

Dupont, W.D., and Plummer, W.D., Jr. (1998). Power and sample size calculations for studies involving linear regression. *Control Clin Trials 19*, 589-601.

Elias-Sonnenschein, L.S., Viechtbauer, W., Ramakers, I.H., Verhey, F.R., and Visser, P.J. (2011). Predictive value of APOE-epsilon4 allele for progression from MCI to AD-type dementia: a meta-analysis. *J Neurol Neurosurg Psychiatry 82*, 1149-1156.

Feldman, H.H., and Jacova, C. (2005). Mild cognitive impairment. *Am J Geriatr Psychiatry 13*, 645-655.

Feldman, H.H., Jacova, C., Robillard, A., Garcia, A., Chow, T., Borrie, M., Schipper, H.M., Blair, M., Kertesz, A., and Chertkow, H. (2008). Diagnosis and treatment of dementia: 2. Diagnosis. *CMAJ 178*, 825-836.

Ge, D., Zhang, K., Need, A.C., Martin, O., Fellay, J., Urban, T.J., Telenti, A., and Goldstein, D.B. (2008). WGAViewer: software for genomic annotation of whole genome association studies. *Genome Res 18*, 640-643.

Geldmacher, D.S., and Whitehouse, P.J. (1996). Evaluation of dementia. *N Engl J Med 335*, 330-336.

Gorelick, P.B., Scuteri, A., Black, S.E., Decarli, C., Greenberg, S.M., Iadecola, C., Launer, L.J., Laurent, S., Lopez, O.L., Nyenhuis, D., *et al.* (2011). Vascular contributions to cognitive impairment and dementia: a statement for healthcare professionals from the american heart association/american stroke association. *Stroke 42*, 2672-2713.

Graham, J.E., Rockwood, K., Beattie, B.L., Eastwood, R., Gauthier, S., Tuokko, H., and McDowell, I. (1997). Prevalence and severity of cognitive impairment with and without dementia in an elderly population. *Lancet 349*, 1793-1796.

Harbison, S.T., Yamamoto, A.H., Fanara, J.J., Norga, K.K., and Mackay, T.F. (2004). Quantitative trait loci affecting starvation resistance in Drosophila melanogaster. *Genetics 166*, 1807-1823.

Hegele, R.A. (2009). Plasma lipoproteins: genetic influences and clinical implications. *Nat Rev Genet 10*, 109-121.

Hirschhorn, J.N., and Daly, M.J. (2005). Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet 6*, 95-108.

Hollingworth, P., Harold, D., Sims, R., Gerrish, A., Lambert, J.C., Carrasquillo, M.M., Abraham, R., Hamshere, M.L., Pahwa, J.S., Moskvina, V., *et al.* (2011). Common variants at ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP are associated with Alzheimer's disease. *Nat Genet 43*, 429-435.

Holtzman, D.M., Herz, J., and Bu, G. (2012). Apolipoprotein e and apolipoprotein e receptors: normal biology and roles in Alzheimer disease. *Cold Spring Harb Perspect Med 2*, a006312.

Hsiung, G.Y., Sadovnick, A.D., and Feldman, H. (2004). Apolipoprotein E epsilon4 genotype as a risk factor for cognitive decline and dementia: data from the Canadian Study of Health and Aging. *CMAJ 171*, 863-867.

Huang, Y., and Mucke, L. (2012). Alzheimer mechanisms and therapeutic strategies. *Cell 148*, 1204-1222.

Kalaria, R.N. (2000). The role of cerebral ischemia in Alzheimer's disease. *Neurobiol Aging 21*, 321-330.

Naj, A.C., Jun, G., Beecham, G.W., Wang, L.S., Vardarajan, B.N., Buros, J., Gallins, P.J., Buxbaum, J.D., Jarvik, G.P., Crane, P.K., *et al.* (2011). Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease. *Nat Genet 43*, 436-441.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., *et al.* (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet 81*, 559-575.

Schneider, J.A., Arvanitakis, Z., Bang, W., and Bennett, D.A. (2007). Mixed brain pathologies account for most dementia cases in community-dwelling older persons. *Neurology 69*, 2197-2204.

Schrijvers, E.M., Schurmann, B., Koudstaal, P.J., van den Bussche, H., Van Duijn, C.M., Hentschel, F., Heun, R., Hofman, A., Jessen, F., Kolsch, H., *et al.* (2011). Genome-wide association study of vascular dementia. *Stroke 43*, 315-319.

Speliotes, E.K., Willer, C.J., Berndt, S.I., Monda, K.L., Thorleifsson, G., Jackson, A.U., Allen, H.L., Lindgren, C.M., Luan, J., Magi, R., *et al.* (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat Genet 42*, 937-948.

Tarawneh, R., and Holtzman, D.M. (2012). The clinical problem of symptomatic Alzheimer disease and mild cognitive impairment. *Cold Spring Harb Perspect Med 2*, a006148.

Teng, E.L., and Chui, H.C. (1987). The Modified Mini-Mental State (3MS) examination. *J Clin Psychiatry 48*, 314-318.

Teslovich, T.M., Musunuru, K., Smith, A.V., Edmondson, A.C., Stylianou, I.M., Koseki, M., Pirruccello, J.P., Ripatti, S., Chasman, D.I., Willer, C.J., *et al.* (2010). Biological, clinical and population relevance of 95 loci for blood lipids. *Nature 466*, 707-713.

Tuokko, H., Frerichs, R., Graham, J., Rockwood, K., Kristjansson, B., Fisk, J., Bergman, H., Kozma, A., and McDowell, I. (2003). Five-year follow-up of cognitive impairment with no dementia. *Arch Neurol 60*, 577-582.

Tuokko, H.A., Frerichs, R.J., and Kristjansson, B. (2001). Cognitive impairment, no dementia: concepts and issues. *Int Psychogeriatr 13 Supp 1*, 183-202.

Wagner, S.A., Beli, P., Weinert, B.T., Nielsen, M.L., Cox, J., Mann, M., and Choudhary, C. (2011). A proteome-wide, quantitative survey of in vivo ubiquitylation sites reveals widespread regulatory roles. *Mol Cell Proteomics 10*, M111.013284.

Wolf, P.A. (2012). Contributions of the Framingham Heart Study to Stroke and Dementia Epidemiologic Research at 60 Years. *Arch Neurol 69*, 567-571.

CHAPTER 4

INVESTIGATING TYPE 2 DIABETES-ASSOCIATED COMMON VARIATION IN ABORIGINAL POPULATIONS

4.1 INTRODUCTION

Globally, aboriginal and indigenous communities are facing an escalating type 2 diabetes (T2D) epidemic. T2D was virtually non-existent amongst aboriginal populations decades ago (Chase, 1937). However, aboriginal populations such as Ontario Oji-Cree now report T2D in adults at an age-adjusted prevalence of ~26%, which ranks among the highest in the world; aboriginal populations in Arizona, Oklahoma and the Dakotas reported notably higher T2D frequencies of 38%, 40% and 40%, respectively (Yu and Zinman, 2007). Adult Metis in Ontario have a reported diabetes frequency of 11%, which represents a ~25% increase above the national rate (Shah et al., 2011). Even the Canadian Inuit, who historically were untouched by T2D, have now matched national standards for T2D frequency based on a recent study of 36 Canadian Inuit communities (Egeland et al., 2011). This significant and recent expansion of T2D prevalence in virtually all Canadian aboriginal communities suggests that these populations may be at greater risk of T2D compared to the non-aboriginal population.

Explanations for the recent rise in T2D cases among aboriginal people centre on the significant lifestyle changes that Canadian aboriginals have gradually adopted. In comparison to the traditional nomadic hunter-gatherer lifestyles of Canadian aboriginal

groups, the pervasion of Western culture into aboriginal communities has increasingly integrated non-traditional diets and lifestyles. Sedentary occupations and leisure lifestyles, together with diets that have become increasingly high in sugar and fat, have certainly contributed to the stark increase in T2D experienced by aboriginal Canadians and this is reflected in the rising incidence of T2D particularly amongst aboriginal youth (Millar and Dean, 2012; Sellers et al., 2009). Accordingly, nutritional and lifestyle management strategies have been the primary initiatives in combating T2D in aboriginal communities. Although an increase in T2D among aboriginal Canadians may not be unexpected, given the relatively recent introduction of Western diets in aboriginal communities, the rate at which T2D has exploded among Canadian aboriginals is remarkable and has suggested that environmental factors are not entirely responsible for the T2D epidemic. Thus prevailing hypotheses have implicated the role of genetic factors in additionally modulating T2D susceptibility among aboriginal Canadians.

Indeed genetic variation has been associated with T2D susceptibility in aboriginal Canadians. Hegele *et al*. identified a private, common variant in the hepatocyte nuclear factor 1 homeobox A (*HNF1A*) gene in Ontario Oji-Cree descendants. Sequencing of *HNF1A*, which encodes a transcription factor that regulates expression of several liver-specific genes, revealed a non-synonymous mutation of glycine to serine at amino acid 319 (p.G319S) (Hegele et al., 1999a). Almost 40% of adult T2D Oji-Cree patients studied carried the p.G319S variant and these individuals were at increased risk of diabetes, particularly T2D, as p.G319S lowered the T2D age at onset following a gene dosage effect. Subsequent *in vitro* functional analyses of the p.G319S variant suggested

lowered insulin secretion due to a combination of reduced *HNF1A* mRNA expression via the introduction of alternative splicing events as well as reduced HNF1A transactivation activity (Bjorkhaug et al., 2005; Harries et al., 2008). The p.G319S variant provided evidence of a private T2D susceptibility variant among the Oji-Cree and spurred the search for additional genetic variants that may also contribute to T2D risk in the Oji-Cree and indeed other aboriginal populations.

Additionally, Hegele *et al*. used a genome-wide scan using 190 microsatellite markers to agnostically investigate additional T2D susceptibility loci in the Oji-Cree (Hegele et al., 1999b). Although this initial scan revealed potential T2D susceptibility loci, the study was limited in terms of genome coverage and resolution. More recently, genome-wide association studies (GWAS) have been used to identify T2D-related loci by testing for genetic association using markers of common variation known as single nucleotide polymorphisms (SNPs) that occur approximately every 300 nucleotides across the genome. To date, the largest T2D GWAS have involved cohorts of European and Asian descent (Kooner et al., 2011; Voight et al., 2010). Testing for association between T2D and millions of SNPs in multi-ethnic populations has identified several T2D-associated loci which have implicated pathways related to β-cell dysfunction and insulin secretion in modulating T2D risk (Voight et al., 2010). Additional GWAS on glycemic traits have confirmed overlap between at least five T2D susceptibility loci and fasting blood glucose (FBG)-associated loci suggesting shared biological mechanism between disease status and disease-related clinical trait (Billings and Florez, 2010). Surprisingly, relatively few genetic studies have been performed within aboriginal communities despite

global aboriginal populations showing remarkable increases in T2D frequency (Yu and Zinman, 2007).

Using T2D GWAS meta-analyses from European and South Asian cohorts, we sought to investigate the frequencies of established T2D-associated variants in a subset of Canadian aboriginal populations. Our primary objective was to test for association between established T2D-associated variants and T2D status in 2 Canadian aboriginal populations; the Oji-Cree of Sandy Lake, Ontario and the Inuit of Inuvik, Northwest Territories. We also tested for association between T2D variants and FBG in Sandy Lake, Inuvik and Greenland non-diabetic aboriginal populations. Our secondary objective was to compare the accumulation of T2D variants in T2D patients and controls from the Sandy Lake and Inuvik populations in the form of a composite genetic risk score (GRS). We demonstrate that the established T2D-associated variants are not strongly associated with T2D in either of these Canadian aboriginal populations individually or as part of a risk score. However, the Inuvik Inuit T2D patients showed a significantly higher risk score based on South Asian-identified T2D variants compared to non-diabetic Inuvik controls. Furthermore, a South Asian-identified T2D-associated variant in the gene encoding high mobility group 20A (*HMG20A*) was significantly associated with FBG in a combined cohort of Inuvik and Greenland Inuit suggesting a potential T2D susceptibility locus common between South Asians and Inuit.

4.2 MATERIALS AND METHODS

4.2.1 Study populations

All individuals involved in this study gave informed consent and the use of the DNA samples given for research was approved by the University of Western Ontario Research Ethics Board for Health Sciences Research Involving Human Subjects (Review number 07920E, Appendix A-1). Study participants were randomly sampled from 2 Canadian aboriginal groups and 1 Greenland Inuit group. Relevant demographic data are listed in **Table 4.1**. Participants of Oji-Cree descent from Sandy Lake (n=399), Ontario were randomly selected from the Sandy Lake Health and Diabetes Project (Harris et al., 1997). Fasting blood samples were collected at the time of the health study in order to measure plasma glucose as well as for DNA extraction. Participants of Inuit descent (n=282) were randomly selected from the Inuvik Inuit community of the Northwest Territories as well as Greenland and Denmark (n=187) (Bjerregaard et al., 2003).

4.2.2 Study design

Using clinical diagnoses of T2D, we performed tests for association between a panel of 17 T2D GWAS-identified SNPs and T2D status (**Table 4.2**). We performed this study with Sandy Lake T2D cases (n=68) and controls (n=320) in addition to Inuvik T2D cases (n=13) and controls (n=247). FBG concentrations were also available for the Sandy Lake (n=321), Inuvik (n=136), and Greenland (n=187) populations. Within each population, we tested for association between blood glucose concentration and T2D-associated SNP genotypes. Whole-genome amplified DNA samples from each aboriginal group were

Table 4.1. Canadian aboriginal study population demographics.

	Sandy Lake	Inuvik	Greenland
n	399	282	187
Male (%)	44	33	40
Diabetic (%)	17.5	4.6	n.d.
Age (years)	29±16	45±16	43±15
BMI	27±6	31±7	26±5
TG (mmol/L)	1.4±0.7	1.74±1.3	1.0±0.5
Fasting glucose (mmol/L)	6.3±2.9	5.3±0.8	5.7±0.9

BMI, body mass index; n.d., no data; TG, triglyceride concentration.

Table 4.2 Top T2D-associated variants identified by GWAS in European and South Asian cohorts.

European-identified T2D variants

CHR	SNP	Nearest Gene	Position	Alleles (min/maj)	RAF		OR	P-value	Reference
					EU	SA			
3	rs4607103	*ADAMTS9*	64,711,904	T/C	0.76	nr	1.09	$1.2x10^{-8}$	(Zeggini et al., 2008)
3	rs1801282	*PPARG*	12,393,125	G/C	0.86	nr	1.14	$1.7x10^{-6}$	(Saxena et al., 2010)
3	rs4402960	*IGF2BP2*	185,511,687	T/G	0.29	nr	1.14	$8.9x10^{-16}$	
6	rs7754840	*CDKAL1*	20,661,250	C/G	0.31	nr	1.12	$4.1x10^{-11}$	
7	rs864745	*JAZF1*	28,180,556	C/T	0.50	nr	1.10	$5.0x10^{-14}$	
8	rs13266634	*SLC30A8*	118,184,783	T/C	0.65	0.76	1.12	$5.3x10^{-8}$	
9	rs10811661	*CDKN2A/2B*	22,134,094	C/T	0.83	nr	1.20	$7.8x10^{-15}$	
10	rs1111875	*HHEX*	94,462,882	T/C	0.53	nr	1.13	$5.7x10^{-10}$	
10	rs7903146	*TCF7L2*	114,758,349	T/C	0.26	0.30	1.37	$1.0x10^{-48}$	
11	rs5219	*KCNJ11*	17,409,572	T/C	0.47	nr	1.14	$6.7x10^{-11}$	
12	rs7961581	*TSPAN8*	71,663,102	C/T	0.27	nr	1.09	$1.1x10^{-9}$	

South Asian-identified T2D variants

CHR	SNP	Nearest Gene	Position	Alleles (min/maj)	RAF		OR	P-value	Reference
					EU	SA			
2	rs3923113	*GRB14*	165,210,095	C/A	0.64	0.74	1.08	$1.6x10^{-9}$	(Kooner et al., 2011)
3	rs16861329	*ST6GAL1*	188,149,155	A/G	0.86	0.75	1.08	$1.3x10^{-7}$	
10	rs1802295	*VPS26A*	70,601,480	A/G	0.31	0.26	1.07	$2.1x10^{-8}$	
15	rs7178572	*HMG20A*	75,534,245	A/G	0.71	0.52	1.08	$9.2x10^{-13}$	
15	rs2028299	*AP3S2*	88,175,261	C/A	0.31	0.31	1.08	$1.2x10^{-11}$	
20	rs4812829	*HNF4A*	42,422,681	A/G	0.19	0.29	1.09	$8.2x10^{-12}$	

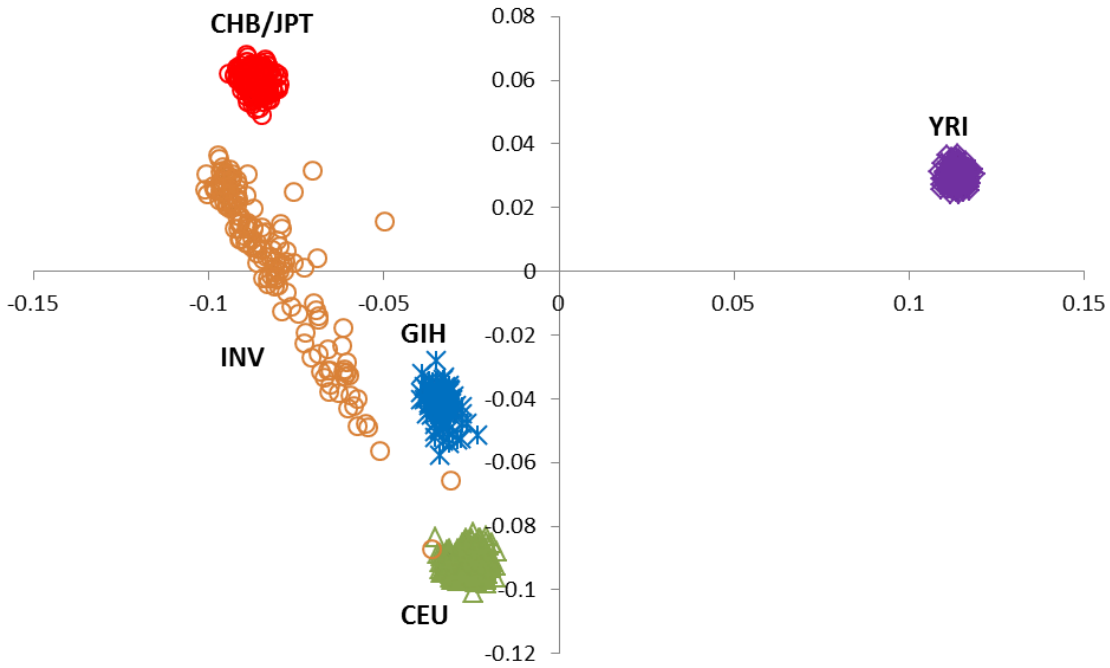
CHR, chromosome; EU, European; maj, major or more frequent allele; min, minor or less frequent allele; OR, odds ratio; RAF, risk allele frequency; SA, South Asian. Risk alleles are underlined.

Variants included in our analysis had call-rates >90%, minor allele frequency (MAF) >1% and were in Hardy Weinberg equilibrium (P>0.05). The 17 T2D SNPs were cumulatively assessed as a composite GRS in order to assess the combined effect of these variants on T2D susceptibility as well as FBG concentration.

4.2.3 Statistical analyses

Study cohort demographics were evaluated using chi-square ($\chi 2$) tests for dichotomous variables and t-tests for continuous variables using SAS v9.2 (Cary, NC); the nominal level of statistical significance was set at P<0.05. Logistic regression was used to test for association between T2D variants and T2D status using PLINK (http://pngu.mgh.harvard.edu/purcell/plink/). Statistical significance was defined as a Bonferroni-corrected P<0.0029. Our logistic regression model was adjusted for potentially confounding variables including age, sex, body-mass index (BMI) and aboriginal population status in analyses combining populations. Similarly, our linear regression model was adjusted for age, sex, BMI and aboriginal population status when combining aboriginal populations. ~200,000 SNP genotypes were available for a subset of Inuvik Inuit (n=142) from genotyping on the custom Cardio-Metabochip genotyping array (Illumina Inc.; San Diego, CA) which were used to perform identity-by-state and multidimensional scaling experiments within this specific subset of Inuvik Inuit (**Figure 4.1**). Genetic risk scores were constructed using study participants with complete genotyping for the 17 T2D-associated SNPs. As each SNP includes an allele associated with T2D risk, each participant's risk score reflects the sum of T2D-associated risk alleles for each of the 17 SNPs. Mean risk scores were compared between T2D patients

Figure 4.1 Principal components analysis with Inuvik and HapMap-derived populations. The first two principal components were plotted with HapMap populations of known ancestry to compare the reported ancestry of Inuvik participants. CEU (green, n=267), Caucasians; INV (orange, n=146), Inuvik; CHB/JPT (red, n=250), Chinese/Japanese; GIH, (blue, n=101), East Indian; YRI, (purple, n=203), African.


CHB/JPT
YRI
INV
GIH
CEU
0.08
0.06
0.04
0.02
0
-0.02
-0.04
-0.06
-0.08
-0.1
-0.12
-0.15
-0.1
-0.05
0
0.05
0.1
0.15

and healthy controls using t-tests. Risk scores for non-diabetics were also compared with FBG concentrations by creating risk score bins and calculating mean FBG concentrations per risk score bin. The strength of correlation between risk score and FBG concentrations was judged based on the r^2 value for the line of best fit where statistical significance was defined by $P<0.05$.

4.2.4 Power calculations

In order to have at least 80% power to detect a true positive association for a SNP representative of our SNP panel (MAF=0.20, OR=1.09, P=0.05), our study required approximately 8000 T2D patients and 32,000 healthy controls. Relative to these calculations, our study cohort was considerably underpowered. However, no Aboriginal-based populations or databases currently exist that would meet these requirements for adequate statistical power. Power calculations were performed using Power and Sample Size Calculation software (Dupont and Plummer, 1990).

4.3 RESULTS

4.3.1 Study participants

Demographic data for each aboriginal group tested for association is shown in **Table 4.1**. T2D status was only available for the Sandy Lake Oji Cree and Inuvik Inuit. Based on these population samples, the Sandy Lake Oji Cree sample showed the highest frequency of T2D participants at 17.5%, which recapitulated the notoriously high T2D prevalence reported in the greater Sandy Lake aboriginal population; mean FBG was also highest

among the Sandy Lake Oji Cree. T2D frequency within the Inuvik Inuit sample was comparatively low, which was representative of the low T2D frequency observed in the greater Inuvik Inuit population.

4.3.2 Establishing T2D variant frequencies in aboriginal populations

First, we sought to determine allele frequencies for each of our candidate T2D variants in our aboriginal study populations. As variant frequencies are known to differ across multi-ethnic populations, we were not surprised to observe variance in T2D risk-associated allele frequencies between the three aboriginal populations and the reference European and South Asian populations (**Table 4.3**). While the majority of variants showed similar allele frequencies across all study and reference populations, it was interesting to observe considerable decreases in risk allele frequencies for variants near *TSPAN8* and *AP3S2* when comparing the European and South Asian cohorts versus our aboriginal study populations. Furthermore, it was also noteworthy that the *HNF4A* risk allele represented the minor allele in the reference populations but approached or attained major allele status within our aboriginal study populations.

4.3.3 Replication of T2D variant associations in aboriginal populations

First, we tested for association between 17 T2D-associated SNPs and T2D status in the Sandy Lake Oji Cree and Inuvik Inuit samples. T2D patients from both aboriginal groups showed significantly higher BMI, TG and fasting glucose compared to controls as expected. T2D patients were also, on average, significantly older than controls (**Table 4.4**). Our logistic regression analysis failed to replicate association that approached

Table 4.3 Putative T2D risk allele frequencies in select aboriginal populations.

European-identified T2D variants								
CHR	**SNP**	**Nearest Gene**	**Alleles (min/maj)**	**RAF**		**RAF**		
				EW	**SA**	**SL**	**IN**	**GR**
3	rs4607103	*ADAMTS9*	T/C	0.76	nr	0.62	0.62	0.76
3	rs1801282	*PPARG*	G/C	0.86	nr	0.92	0.82	0.86
3	rs4402960	*IGF2BP2*	T/G	0.29	nr	0.20	0.26	0.25
6	rs7754840	*CDKAL1*	C/G	0.31	nr	0.36	0.41	0.48
7	rs864745	*JAZF1*	C/T	0.50	nr	0.67	0.50	0.54
8	rs13266634	*SLC30A8*	T/C	0.65	0.76	0.86	0.55	0.67
9	rs10811661	*CDKN2A/2B*	C/T	0.83	nr	0.91	0.64	0.79
10	rs1111875	*HHEX*	T/C	0.53	nr	0.34	0.42	0.65
10	rs7903146	*TCF7L2*	T/C	0.26	0.30	0.14	0.11	0.11
11	rs5219	*KCNJ11*	T/C	0.47	nr	0.16	0.27	0.49
12	rs7961581	*TSPAN8*	C/T	0.27	nr	0.01	0.05	0.13
South Asian-identified T2D variants								
CHR	**SNP**	**Nearest Gene**	**Alleles (min/maj)**	**RAF**		**RAF**		
				EW	**SA**	**SL**	**IN**	**GR**
2	rs3923113	*GRB14*	C/A	0.64	0.74	0.95	0.85	0.84
3	rs16861329	*ST6GAL1*	A/G	0.86	0.75	0.43	0.77	0.78
10	rs1802295	*VPS26A*	A/G	0.31	0.26	0.15	0.63	0.37
15	rs7178572	*HMG20A*	A/G	0.71	0.52	0.40	0.56	0.69
15	rs2028299	*AP3S2*	C/A	0.31	0.31	0.02	0.15	0.14
20	rs4812829	*HNF4A*	A/G	0.19	0.29	0.78	0.43	0.55

Abbreviations as in Table 4.2. SL, Sandy Lake; IN, Inuvik; and GR, Greenland.

Table 4.4 Demographics for T2D and non-T2D patients in two aboriginal Canadian populations.

	Sandy Lake			Inuvialuit		
	T2D	**Non-T2D**	**P-value**	**T2D**	**Non-T2D**	**P-value**
n	68	320	n.a.	13	247	n.a.
% Male	42	45	n.s.	31	32	n.s.
Age (years)	45±15	26±14	<0.0001	65±8	43±15	<0.0001
BMI	31±5	26±6	<0.0001	33±5	30±7	n.s.
TG (mmol/L)	2.1±0.9	1.26±0.62	<0.0001	2.8±1.9	1.7±1.2	0.0035
Fasting glucose (mmol/L)	10.8±4.7	5.41±0.55	<0.0001	6.9±1.7	5.2±0.6	<0.0001

P-values are based on t-tests performed between T2D and non-T2D patients.

nominal significance in either of the Sandy Lake Oji Cree or the Inuvik Inuit cohorts, although some variants, such as those in *PPARG* and *KCNJ11*, replicated similar effect sizes as those previously observed in GWAS meta-analysis (**Table 4.5**).

4.3.4 Association between T2D variants and fasting blood glucose

Next, we tested for association between the 17 T2D-associated SNPs and FBG using an adjusted linear regression model including non-T2D patients from the Sandy Lake Oji Cree, Inuvik Inuit and the Greenland Inuit cohorts (**Table 4.6**). The Sandy Lake Oji Cree cohort showed associations of nominal significance ($P<0.05$) for variants at the *CDKAL1*, *CDKN2A/2B,* and *RPS26A* loci. The Inuvik Inuit cohort showed associations of nominal significance at the *ADAMTS9, TCF7L2, TSPAN6,* and *HMG20A* loci. The Greenland Inuit cohort showed only one association of nominal significance at the *HMG20A* locus. Because the Inuvik and Greenland Inuit share similar ancestry, both cohorts were merged (n=436) and re-analyzed using a linear regression model adjusted for covariates as well as population identity. The combined analysis showed that each copy of the T2D-associated variant in *HMG20A*, also the major allele, was associated with a 0.18 mmol/L increase in FBG ($P=1.6x10^{-4}$). The significance of this association in the combined Inuit cohort surpassed a Bonferroni-corrected level of significance ($P<2.9x10^{-3}$). This association suggests a potential biological connection between a T2D susceptibility locus and FBG concentration in Inuit descendants. Further assessment is required in additional Inuit cohorts to confirm this association.

Table 4.5 Association between established T2D variants and T2D status in aboriginal Canadian populations.

CHR	SNP	Nearest gene	Allele (min/maj)	Ref. OR	Sandy Lake		Inuvialuit	
					OR (95% CI)	P-value	OR (95% CI)	P-value
2	rs3923113	*GRB14*	C/A	1.08	0.53 (0.15 – 1.92)	0.33	2.02E-08 (0 – ∞)	0.99
3	rs1801282	*PPARG*	G/C	1.14	1.11 (0.53 – 2.34)	0.78	0.61 (0.15 – 2.43)	0.49
3	rs4607103	*ADAMTS9*	T/C	1.09	1.13 (0.75 – 1.71)	0.56	1.35 (0.50 – 3.66)	0.55
3	rs4402960	*IGF2BP2*	T/G	1.14	0.84 (0.50 – 1.41)	0.52	0.59 (0.20 – 1.76)	0.35
3	rs16861329	*ST6GAL1*	A/G	1.08	0.84 (0.56 – 1.27)	0.41	1.51 (0.55 – 4.15)	0.43
6	rs7754840	*CDKAL1*	C/G	1.12	1.01 (0.66 – 1.55)	0.96	0.65 (0.24 – 1.75)	0.40
7	rs864745	*JAZF1*	C/T	1.10	0.90 (0.57 – 1.42)	0.65	1.47 (0.55 – 3.94)	0.44
8	rs13266634	*SLC30A8*	T/C	1.12	0.91 (0.49 – 1.71)	0.77	1.73 (0.65 – 4.61)	0.27
9	rs10811661	*CDKN2A/2B*	C/T	1.20	1.46 (0.74 – 2.85)	0.27	0.76 (0.27 – 2.11)	0.60
10	rs1802295	*VPS26A*	A/G	1.07	1.57(0.83 – 2.90)	0.16	0.98 (0.36 – 2.62)	0.96
10	rs1111875	*HHEX*	T/C	1.13	1.40 (0.87 – 2.26)	0.16	1.44 (0.54 – 3.85)	0.46
10	rs7903146	*TCF7L2*	T/C	1.37	0.92 (0.51 – 1.65)	0.77	0.95 (0.25 – 3.65)	0.95
11	rs5219	*KCNJ11*	T/C	1.14	1.36 (0.75 – 2.50)	0.31	1.06 (0.41 – 2.70)	0.91
12	rs7961581	*TSPAN8*	C/T	1.09	2.11E+08 (0 – ∞)	0.99	6.09E+07 (0 – ∞)	0.99
15	rs7178572	*HMG20A*	A/G	1.08	0.72 (0.47 – 1.11)	0.13	0.54 (0.21 – 1.43)	0.23
15	rs2028299	*AP3S2*	C/A	1.08	0.83 (0.20 – 3.44)	0.80	0.62 (0.14 – 2.78)	0.53
20	rs4812829	*HNF4A*	A/G	1.09	0.65 (0.37 – 1.12)	0.12	1.02 (0.38 – 2.78)	0.98

P-values shown are based on established risk allele (underlined) and adjusted for sex, age, and body mass index. Abbreviations as in Table 4.1.

Table 4.6 Association between established T2D variants and fasting blood glucose in three aboriginal populations.

CHR	SNP	Nearest Gene	Allele (min/maj)	Population	β (mmol/L)	SE	P-value
2	rs3923113	*GRB14*	C/A	Sandy Lake	-0.032	0.091	0.72
				Inuvialuit	-0.10	0.076	0.19
				Greenland	0.061	0.12	0.60
				Comb. Inuit	-0.051	0.067	0.45
3	rs4607103	*ADAMTS9*	T/C	Sandy Lake	-0.011	0.039	0.77
				Inuvialuit	-0.12	0.059	0.041
				Greenland	0.050	0.10	0.63
				Comb. Inuit	-0.066	0.054	0.23
3	rs1801282	*PPARG*	G/C	Sandy Lake	-0.027	0.077	0.73
				Inuvialuit	0.13	0.071	0.067
				Greenland	0.080	0.12	0.50
				Comb. Inuit	0.11	0.065	0.089
3	rs4402960	*IGF2BP2*	T/G	Sandy Lake	0.037	0.048	0.44
				Inuvialuit	0.016	0.061	0.80
				Greenland	-0.051	0.096	0.60
				Comb. Inuit	-0.0078	0.055	0.89
3	rs16861329	*ST6GAL1*	A/G	Sandy Lake	-0.020	0.039	0.62
				Inuvialuit	0.041	0.066	0.54
				Greenland	0.045	0.10	0.60
				Comb. Inuit	0.067	0.058	0.25

Table 4.6 continued.

CHR	SNP	Nearest Gene	Allele (maj/min)	Population	β (mmol/L)	SE	P-value
6	rs7754840	*CDKAL1*	C/G	Sandy Lake	0.10	0.039	0.015
				Inuvialuit	-0.033	0.057	0.56
				Greenland	-0.017	0.082	0.84
				Comb. Inuit	-0.040	0.048	0.41
7	rs864745	*JAZF1*	C/T	Sandy Lake	-0.052	0.041	0.21
				Inuvialuit	-0.0025	0.055	0.96
				Greenland	0.094	0.089	0.29
				Comb. Inuit	0.025	0.049	0.61
8	rs13266634	*SLC30A8*	T/C	Sandy Lake	0.030	0.062	0.63
				Inuvialuit	-0.040	0.055	0.47
				Greenland	0.076	0.096	0.43
				Comb. Inuit	0.0055	0.051	0.91
9	rs10811661	*CDKN2A/2B*	C/T	Sandy Lake	0.15	0.069	0.030
				Inuvialuit	-0.0093	0.053	0.86
				Greenland	-0.00037	0.11	0.99
				Comb. Inuit	-0.00043	0.052	0.99
10	rs1111875	*HHEX*	T/C	Sandy Lake	-0.016	0.042	0.71
				Inuvialuit	0.0074	0.056	0.89
				Greenland	0.026	0.087	0.76
				Comb. Inuit	0.033	0.049	0.50

Table 4.6 continued.

CHR	SNP	Nearest Gene	Allele (maj/min)	Population	β (mmol/L)	SE	P-value
10	rs7903146	*TCF7L2*	T/C	Sandy Lake	-0.086	0.058	0.14
				Inuvialuit	0.19	0.081	0.024
				Greenland	-0.073	0.16	0.64
				Comb. Inuit	0.086	0.079	0.28
10	rs1802295	*VPS26A*	A/G	Sandy Lake	-0.10	0.052	0.044
				Inuvialuit	-0.050	0.057	0.38
				Greenland	-0.011	0.095	0.91
				Comb. Inuit	-0.035	0.052	0.50
11	rs5219	*KCNJ11*	T/C	Sandy Lake	-0.016	0.052	0.76
				Inuvialuit	-0.062	0.060	0.30
				Greenland	0.11	0.082	0.17
				Comb. Inuit	0.034	0.050	0.50
12	rs7961581	*TSPAN8*	C/T	Sandy Lake	-0.052	0.17	0.76
				Inuvialuit	-0.25	0.12	0.037
				Greenland	-0.039	0.13	0.76
				Comb. Inuit	-0.14	0.088	0.11
15	rs7178572	*HMG20A*	A/G	Sandy Lake	0.053	0.039	0.18
				Inuvialuit	0.13	0.052	0.017
				Greenland	0.26	0.086	2.1×10^{-3}
				Comb. Inuit	0.18	0.048	1.6×10^{-4}

Table 4.6 continued.

CHR	SNP	Nearest Gene	Allele (min/maj)	Population	β (mmol/L)	SE	P-value
15	rs2028299	*AP3S2*	C/A	Sandy Lake	0.052	0.16	0.74
				Inuvialuit	0.085	0.082	0.30
				Greenland	0.10	0.12	0.39
				Comb. Inuit	0.10	0.070	0.15
20	rs4812829	*HNF4A*	A/G	Sandy Lake	0.028	0.049	0.57
				Inuvialuit	0.020	0.056	0.72
				Greenland	-0.072	0.088	0.42
				Comb. Inuit	-0.019	0.050	0.70

Values presented from a linear regression model adjusted for age, sex, body mass index and population identity (combined analysis only). Inuvialuit and Greenland populations were combined and re-analysed. Underlined alleles represent established type 2 diabetes-associated risk alleles. β,effect size; SE, standard error. Comb., combined samples. Remaining abbreviations as in Table 4.1.

4.3.5 T2D genetic risk scores in aboriginal populations

Finally, we tested whether T2D patients from the Sandy Lake Oji Cree and Inuvik Inuit cohorts carried a significantly greater accumulation of multiple T2D-associated variants through the use of a composite GRS. There was no significant difference between mean risk scores using all 17 SNPs for T2D patients and healthy controls in either the Sandy Lake Oji Cree or Inuvik Inuit cohorts. As a sub-analysis, we sub-divided risk scores based on T2D SNPs identified in either the European or South Asian cohort meta-analyses. No significant difference was observed in either the Sandy Lake Oji Cree or Inuvik Inuit cohorts based on the European T2D SNP risk score. Conversely, we observed a nominally significant difference in mean South Asian-derived risk scores between Inuvik Inuit T2D patients and healthy controls (**Table 4.7, Figure 4.2C**). The same South Asian risk score adjusted for the reported effect sizes for each risk allele produced a similar nominally significant result (**Table 4.7**). We further tested whether risk score correlates with FBG independently in combined non-diabetic Inuvik (n=184) and Greenlanders of unknown diabetic status (n=159) using regression analysis; however no association was identified (**Figure 4.3**).

4.4 DISCUSSION

The principal finding of our study is that established T2D-associated variants are not associated with T2D status in Canadian aboriginal patients. We were unable to replicate associations between 17 GWAS-identified T2D variants and the Sandy Lake Oji-Cree or

Table 4.7 Genetic risk scores in T2D and non-T2D patients in two aboriginal Canadian populations.

Non-weighted				
	GRS type	**Non-T2D GRS**	**T2D GRS**	**P-value**
Sandy Lake	**European**	10.4±0.1	10.1±0.3	0.29
	South Asian	4.4±0.1	4.3±0.2	0.74
	Cumulative	14.8±0.1	14.4±0.3	0.28
Inuvik Inuit	**European**	11.2±0.1	10.8±0.6	0.48
	South Asian	7.2±0.1	8.1±0.3	0.049
	Cumulative	18.4±0.2	18.9±0.6	0.51
Weighted				
	GRS type	**Non-T2D GRS**	**T2D GRS**	**P-value**
Sandy Lake	**European**	11.7±0.1	11.4±0.3	0.29
	South Asian	4.8±0.09	4.7±0.2	0.74
	Cumulative	16.5±0.2	16.1±0.4	0.28
Inuvik Inuit	**European**	12.4±0.2	12.0±0.7	0.61
	South Asian	7.8±0.1	8.8±0.4	0.049
	Cumulative	20.5±0.2	20.9±0.7	0.58

Mean risk scores are shown ± standard deviation in cohorts of T2D patients and healthy controls. GRS, genetic risk score.

Figure 4.2. Frequency distributions of non-weighted T2D risk scores in Inuvik T2D patients and healthy controls. Inuvik T2D patients (n=11) and healthy Inuvik controls (n=184) with complete genotyping were included. A) Combined risk scores were calculated using a total of 17 T2D variants identified in both European and South Asian cohorts (min. score=0, max. score=34). Mean risk scores between T2D patients and controls, 18.9 and 18.4 respectively, were non-significant (P=0.51). B) European-derived risk scores included 11 T2D variants identified in T2D GWAS in European cohorts (min. score=0, max. score=22). Mean risk scores between T2D patients and controls, 10.8 and 11.2 respectively, were non-significant (P=0.48). C) South Asian-derived risk scores included 6 T2D variants identified by T2D GWAS in South Asian cohorts (min. score=0, max. score=12). Mean risk scores between T2D patients and controls, 8.1 and 7.2 respectively, were non-significant (P=0.05). Statistics were calculated using Student's t-test.

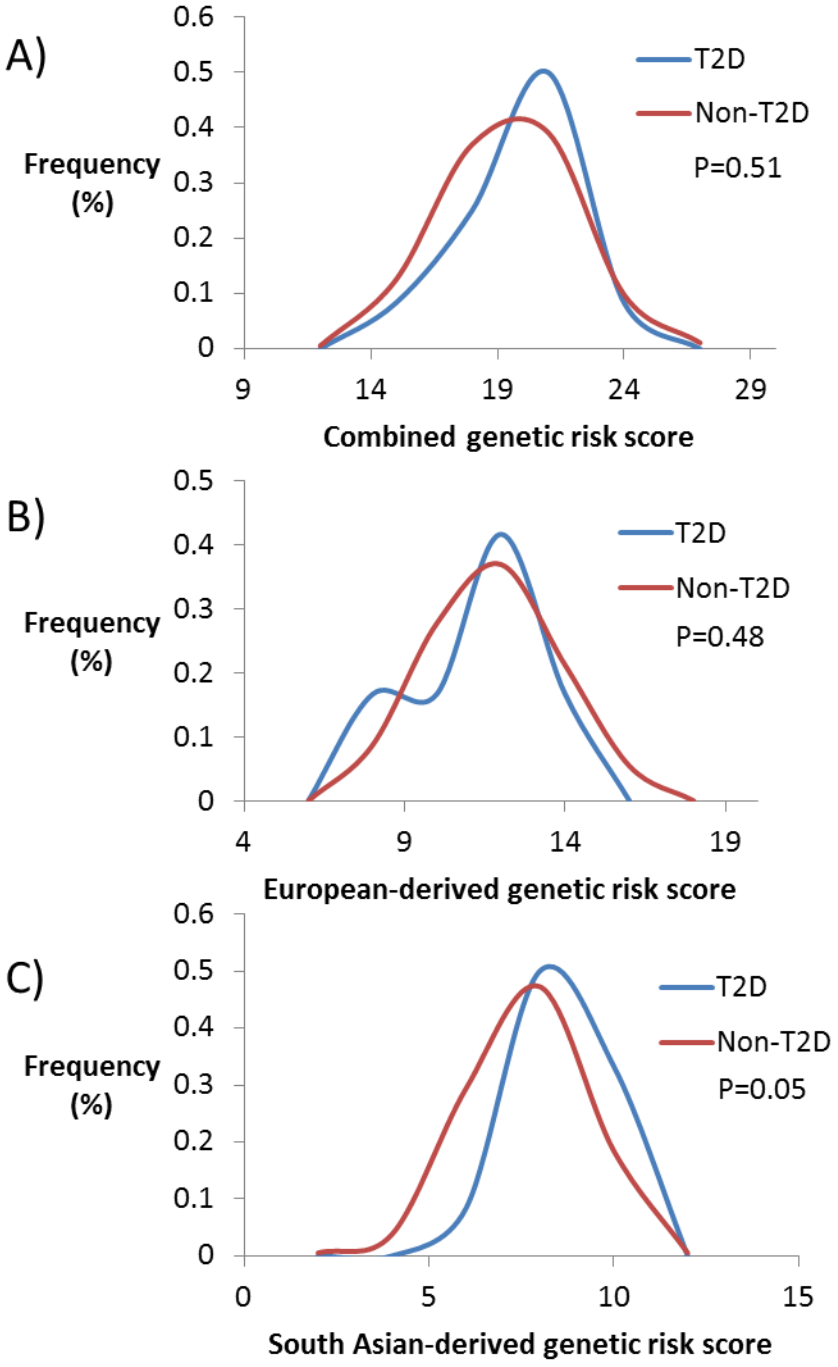
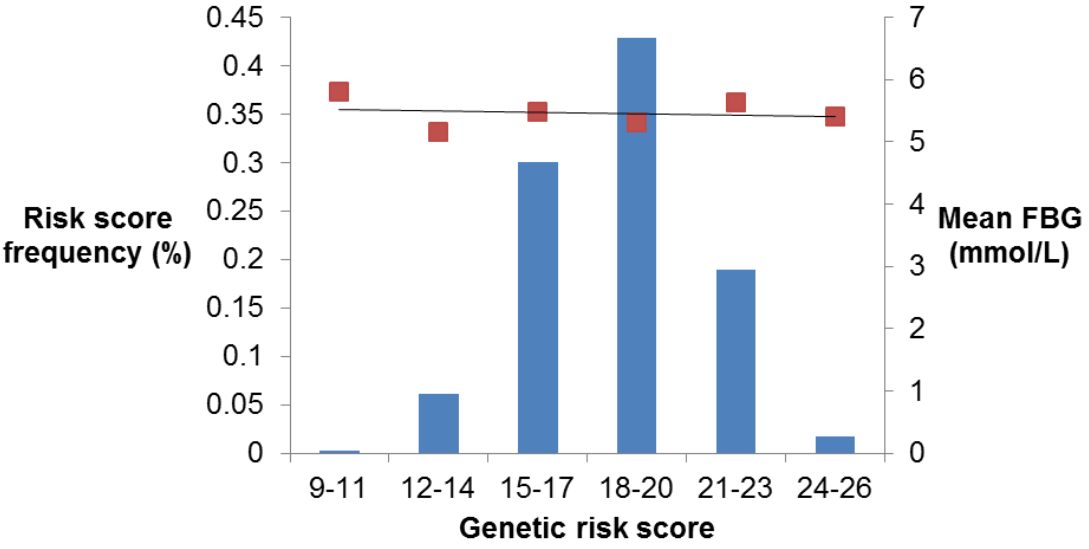

A)
Frequency (%)
0.6
0.5
0.4
0.3
0.2
0.1
0
9
14
19
24
29
Combined genetic risk score
T2D
Non-T2D
P=0.51
B)
Frequency (%)
0.5
0.4
0.3
0.2
0.1
0
4
9
14
19
European-derived genetic risk score
T2D
Non-T2D
P=0.48
C)
Frequency (%)
0.6
0.5
0.4
0.3
0.2
0.1
0
0
5
10
15
South Asian-derived genetic risk score
T2D
Non-T2D
P=0.05

Figure 4.3. Correlation between fasting blood glucose and T2D genetic risk score. Mean fasting blood glucose (FBG) values are plotted for participants sorted into bins of genetic risk score based on 17 T2D variants. Non-diabetics from Inuvik Inuit (n=184) and Greenland Inuit (n=159) populations with complete genotyping for 17 T2D variants were included. However, no significant association was observed (P=0.45). Statistics were calculated from a linear regression model.


0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
Risk score frequency (%)
7
6
5
4
3
2
1
0
Mean FBG (mmol/L)
9-11
12-14
15-17
18-20
21-23
24-26
Genetic risk score

Inuvik Inuit T2D patients. Our T2D GRS analysis supported this finding, however, Inuvik Inuit T2D patients showed a modestly higher accumulation of South Asian T2D risk alleles compared to controls. One significant association was detected between a T2D-associated variant from the *HMG20A* locus (rs7178572) and FBG in the two Inuit populations studied here which suggests a potential T2D-related susceptibility locus common between South Asians and Inuit descendants.

Our study provides novel insight into T2D susceptibility amongst aboriginal Canadians. The largest GWAS of T2D loci to date have focused on cohorts of European, East Asian and South Asian descent and have successfully identified several T2D susceptibility loci of modest effect. Previous genetic studies in aboriginal Canadian populations, guided by a candidate gene approach, identified variants associated with T2D risk and so we sought to apply a similar approach by investigating the top common variants from T2D GWAS in aboriginal Canadian populations. Using a case-control approach in Sandy Lake Oji-Cree and Inuvik Inuit samples, we were unable to replicate previous GWAS-identified associations with T2D. This was not entirely unexpected given statistical limitations due to small sample size involved as well as issues surrounding the replicability of GWAS-identified variants in multi-ethnic populations. The prevailing concept regarding GWAS findings suggests that the top variant at a locus tags additional variants through linkage disequilibrium. Between multi-ethnic populations, however, patterns of linkage disequilibrium differ and so GWAS-identified risk variants may vary from population to population as variant frequencies also differ (Cooper et al., 2008; Fu et al., 2011). Thus, it remains possible that the top T2D loci from

European and South Asian cohorts may not directly translate to T2D susceptibility in Canadian aboriginal populations.

In addition to the dichotomous analysis of T2D status, GWAS have investigated the role of common variation in modulating clinical T2D-related traits such as FBG. These studies, performed in participants without diabetes, were initiated to develop better understanding of the biological pathways involved in regulating diabetes-related quantitative traits. Cumulatively, common variation explains ~10% of the inherited variation in FBG concentration suggesting a modest role for common variation in FBG variability (Dupuis et al., 2010). As several loci have been associated with both T2D status and FBG, such as *GCKR, MTNR1B* and *TCF7L2*, we tested whether our panel of top T2D-associated variants was associated with FBG in three aboriginal populations. Although our findings were largely negative, we identified a modest association between the South Asian-identified *HMG20A* locus (rs7178572) and FBG. In a combined cohort of Inuvik Inuit and Greenland Inuit, we observed that the T2D risk allele for this variant was also associated with a 0.18 mmol/L increase in FBG per copy of the T2D risk allele ($P=1.6x10^{-4}$). As this trend was not observed in the Sandy Lake Oji-Cree sample, our findings suggest that this variant may be unique to South Asian and Inuit descendants however, further validation is required. The observed effect of the *HMG20A* variant on FBG is almost twice that reported by the top GWAS-identified loci on FBG (β= ~0.07 mmol/L) although the clinical relevance of such an effect is not clear.

Although our study is the first to test for association between established T2D variants and T2D status as well as FBG concentration, our findings must be considered within the context of specific limitations. First, our statistical power to detect associations between the 17 T2D-associated variants and T2D was limited by sample size. The T2D variants featured in our study were identified in cohorts that included thousands of participants which is essential to detecting significant associations of small effect size. Due to the limited sample size of the aboriginal populations featured in our study, we cannot rule out the potential for these 17 T2D variants in playing a role in T2D susceptibility among Canadian aboriginals. Furthermore, the modest association detected between the *HMG20A* variant and FBG, while notable, also requires additional validation given our limited statistical power.

Secondly, our approach which targeted GWAS-identified T2D loci is likely to have excluded additional loci that may be uniquely associated with Canadian aboriginal populations. The application of GWAS to a Canadian aboriginal population would facilitate an agnostic search for T2D susceptibility loci, however, the sticking point remains identifying a sizable population to afford adequate statistical power. One option that may help in addressing this issue may be combining aboriginal populations with related ancestry such as the circumpolar Inuit populations.

Thirdly, previous studies have shown that there is often difficulty in replicating GWAS findings across multi-ethnic populations (Cooper et al., 2008; Fu et al., 2011; Imamura and Maeda, 2011). On an individual variant basis, this is partly due to the

variability in allele frequencies that exist between populations of different ethnicity. For example, the lead SNP from the *UBE2E2* locus (rs6780569) has a reported minor allele frequency (MAF) of ~9% in Europeans, however, the same variant has a MAF of ~22% in East Asians (Fu et al., 2011). With a lower MAF, larger sample sizes are required in order to maintain adequate statistical power thus variability in allele frequency may confound the ability to replicate an association in ethnically diverse populations. Accordingly, the lead SNP from *UBE2E2* was associated with T2D in East Asians ($P=1.0x10^{-9}$) but was not replicable in Europeans (P=0.98) (Yamauchi et al., 2010). Although statistical power limitations in our study prevent the ruling out of associations between top T2D variants and aboriginal T2D susceptibility, it may be expected that the established T2D variants may vary in effect size or frequency in Canadian aboriginal populations.

Lastly, limited data on our participants' relatedness to other participants within each aboriginal sample prevented adjustment for any inflation in association signal due to common ancestry. The inclusion of multiple closely related participants in either cases or controls may inflate allele frequencies simply due to relatedness and not due to T2D susceptibility. This effect may also confound quantitative trait analysis as closely related participants may cluster at either end of a quantitative trait spectrum. Future studies must carefully document family structures and relatedness within aboriginal populations especially as these populations are likely to be small with low net migration into these communities.

In summary, we have demonstrated that the top GWAS-identified T2D-associated variants were not associated with T2D susceptibility in a sample of Canadian aboriginals. We also demonstrated, through a composite GRS that aboriginal T2D patients did not carry a significantly higher burden of T2D-associated variants compared to controls. One variant in *HMG20A* was associated with FBG in aboriginals of Inuit descent; however, this observation requires further replication in larger cohorts. Due to the power limitations of our study, the findings reported here have not ruled out the potential role for genetic variation at these established loci in modulating T2D susceptibility and glycemic traits in aboriginal populations. Further studies involving larger aboriginal cohorts with well-documented information on relatedness will contribute towards the ongoing investigation of T2D determinants amongst Canadian aboriginal populations and will provide a more conclusive assessment of the degree of overlap between European- and South Asian-identified T2D variants in aboriginal populations.

4.5 REFERENCES

Billings, L.K., and Florez, J.C. (2010). The genetics of type 2 diabetes: what have we learned from GWAS? *Ann N Y Acad Sci 1212*, 59-77.

Bjerregaard, P., Curtis, T., Borch-Johnsen, K., Mulvad, G., Becker, U., Andersen, S., and Backer, V. (2003). Inuit health in Greenland: a population survey of life style and disease in Greenland and among Inuit living in Denmark. *Int J Circumpolar Health 62 Suppl 1*, 3-79.

Bjorkhaug, L., Bratland, A., Njolstad, P.R., and Molven, A. (2005). Functional dissection of the HNF-1alpha transcription factor: a study on nuclear localization and transcriptional activation. *DNA Cell Biol 24*, 661-669.

Chase, L.A. (1937). The Trend of Diabetes in Saskatchewan, 1905 to 1934. *Can Med Assoc J 36*, 366-369.

Cooper, R.S., Tayo, B., and Zhu, X. (2008). Genome-wide association studies: implications for multiethnic samples. *Hum Mol Genet 17*, R151-155.

Dupont, W.D., and Plummer, W.D., Jr. (1990). Power and sample size calculations. A review and computer program. *Control Clin Trials 11*, 116-128.

Dupuis, J., Langenberg, C., Prokopenko, I., Saxena, R., Soranzo, N., Jackson, A.U., Wheeler, E., Glazer, N.L., Bouatia-Naji, N., Gloyn, A.L., *et al.* (2010). New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nat Genet 42*, 105-116.

Egeland, G.M., Cao, Z., and Young, T.K. (2011). Hypertriglyceridemic-waist phenotype and glucose intolerance among Canadian Inuit: the International Polar Year Inuit Health Survey for Adults 2007-2008. *CMAJ 183*, E553-558.

Fu, J., Festen, E.A., and Wijmenga, C. (2011). Multi-ethnic studies in complex traits. *Hum Mol Genet 20*, R206-213.

Harries, L.W., Sloman, M.J., Sellers, E.A., Hattersley, A.T., and Ellard, S. (2008). Diabetes susceptibility in the Canadian Oji-Cree population is moderated by abnormal mRNA processing of HNF1A G319S transcripts. *Diabetes 57*, 1978-1982.

Harris, S.B., Gittelsohn, J., Hanley, A., Barnie, A., Wolever, T.M., Gao, J., Logan, A., and Zinman, B. (1997). The prevalence of NIDDM and associated risk factors in native Canadians. *Diabetes Care 20*, 185-187.

Hegele, R.A., Cao, H., Harris, S.B., Hanley, A.J., and Zinman, B. (1999a). The hepatic nuclear factor-1alpha G319S variant is associated with early-onset type 2 diabetes in Canadian Oji-Cree. *J Clin Endocrinol Metab 84*, 1077-1082.

Hegele, R.A., Sun, F., Harris, S.B., Anderson, C., Hanley, A.J., and Zinman, B. (1999b). Genome-wide scanning for type 2 diabetes susceptibility in Canadian Oji-Cree, using 190 microsatellite markers. *J Hum Genet 44*, 10-14.

Imamura, M., and Maeda, S. (2011). Genetics of type 2 diabetes: the GWAS era and future perspectives [Review]. *Endocr J 58*, 723-739.

Kooner, J.S., Saleheen, D., Sim, X., Sehmi, J., Zhang, W., Frossard, P., Been, L.F., Chia, K.S., Dimas, A.S., Hassanali, N., *et al.* (2011). Genome-wide association study in individuals of South Asian ancestry identifies six new type 2 diabetes susceptibility loci. *Nat Genet 43*, 984-989.

Millar, K., and Dean, H.J. (2012). Developmental origins of type 2 diabetes in aboriginal youth in Canada: it is more than diet and exercise. *J Nutr Metab 2012*, 127452.

Saxena, R., Hivert, M.F., Langenberg, C., Tanaka, T., Pankow, J.S., Vollenweider, P., Lyssenko, V., Bouatia-Naji, N., Dupuis, J., Jackson, A.U., *et al.* (2010). Genetic variation in GIPR influences the glucose and insulin responses to an oral glucose challenge. *Nat Genet 42*, 142-148.

Sellers, E.A., Moore, K., and Dean, H.J. (2009). Clinical management of type 2 diabetes in indigenous youth. *Pediatr Clin North Am 56*, 1441-1459.

Shah, B.R., Cauch-Dudek, K., and Pigeau, L. (2011). Diabetes prevalence and care in the Metis population of Ontario, Canada. *Diabetes Care 34*, 2555-2556.

Voight, B.F., Scott, L.J., Steinthorsdottir, V., Morris, A.P., Dina, C., Welch, R.P., Zeggini, E., Huth, C., Aulchenko, Y.S., Thorleifsson, G., *et al.* (2010). Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat Genet 42*, 579-589.

Yamauchi, T., Hara, K., Maeda, S., Yasuda, K., Takahashi, A., Horikoshi, M., Nakamura, M., Fujita, H., Grarup, N., Cauchi, S., *et al.* (2010). A genome-wide association study in the Japanese population identifies susceptibility loci for type 2 diabetes at UBE2E2 and C2CD4A-C2CD4B. *Nat Genet 42*, 864-868.

Yu, C.H., and Zinman, B. (2007). Type 2 diabetes and impaired glucose tolerance in aboriginal populations: a global perspective. *Diabetes Res Clin Pract 78*, 159-170.

Zeggini, E., Scott, L.J., Saxena, R., Voight, B.F., Marchini, J.L., Hu, T., de Bakker, P.I., Abecasis, G.R., Almgren, P., Andersen, G., *et al.* (2008). Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat Genet 40*, 638-645.

CHAPTER 5

DISCUSSIONS AND CONCLUSIONS

5.1 Genetic characterization of complex disease

Over the past decade, our concept of genetic predisposition to complex disease has rapidly gained depth. As a result of the efforts of international consortia such as the Human Genome Project and the International HapMap Project (International Hapmap Consortium, 2003), the map of common genomic variation has enabled a leap forward to high-resolution genome-wide association studies (GWAS). At the time of this thesis work, the GWAS approach was established as a versatile and effective tool for discovering novel disease susceptibility loci and validating candidate loci based on the frequencies of common variants. Combined with targeted resequencing, findings from the GWAS era have provided new perspectives on complex disease susceptibility and have also supported re-evaluation of current concepts of complex disease genetics.

The studies presented within this thesis reflect application of the current established techniques to characterize the genetics of complex disease susceptibility. Using candidate gene resequencing, the GWAS approach and targeted genotyping of GWAS-identified variants, we have demonstrated the application of modern genetic approaches to investigate complex diseases relating largely to cardiovascular disease (CVD). We implemented these techniques via 1) a candidate gene resequencing study in

circumpolar Inuit populations that revealed private, common missense variants within the low-density lipoprotein receptor (*LDLR*) gene; 2) the most comprehensive genomic analysis of a pre-dementia phenotype known as "cognitive impairment, no dementia" (CIND) utilizing cardio-metabolic and Alzheimer disease (AD)-associated variants; and 3) a type 2 diabetes (T2D) candidate variant analysis in North American and Greenland aboriginals based on T2D GWAS meta-analyses. Cumulatively, these studies recapitulate the current molecular genetics techniques and analytical approaches widely utilized in order to assess the role of common variation on disease susceptibility and phenotypic variability.

5.1.1 p.G116S in *LDLR* is associated with LDL-C among the Inuit

We have reported the private, common p.G116S and p.R730W variants in *LDLR* within five circumpolar Inuit populations. Furthermore, we showed that G116S was robustly associated with a large effect on plasma low-density lipoprotein cholesterol (LDL-C) while p.R730W showed a modest non-significant effect on LDL-C (**Table 2.4**). Although our statistical analyses have implicated p.G116S as having a considerable effect on plasma LDL-C concentration, additional biochemical experiments are required in order to establish causal mechanisms underlying the observed p.G116S association with higher LDL-C. As follow-up experiments, we will investigate the effects of either *LDLR* variant on LDLR expression as well as receptor activity.

Our findings regarding p.G116S and p.R730W have provided new insight into the unique genetic architecture of Inuit descendants as well as potential CVD risk factors

exclusive to the Inuit. As LDLR has been well-established in cholesterol homeostasis, we applied direct Sanger sequencing of *LDLR* coding regions to test for the presence of coding variants associated with plasma LDL-C. Interestingly, the variants p.G116S and p.R730W were observed with relatively high frequencies (5%-17%) across five distinct Inuit samples (**Table 2.1, Figure 2.1**). In line with the fact that LDLR is a major regulator of LDL-C homeostasis, we replicated independent associations between mean LDL-C and p.G116S or p.R730W carrier status in a combined Inuit cohort (**Table 2.3A**). These findings were further explored by using multivariate linear regression to test a dominant genetic model based on either p.G116S or p.R730W status which further supported association between p.G116S and LDL-C (**Table 2.4**). Importantly, p.G116S was consistently associated with raising LDL-C within each Inuit population with a summary effect size of ~0.54 mmol/L per allele dose in a combined Inuit cohort while p.R730W was non-significantly linked with a modest lowering of LDL-C by ~0.05 mmol/L per allele dose (**Table 2.4**). The effect of p.G116S on LDL-C is intriguing as common variants are not usually associated with such large effect sizes. The top LDL-C-associated common variants from GWAS meta-analyses near *SORT1, APOE* and *LDLR* were reported with per-allele effect sizes of 0.15 mmol/L, 0.18 mmol/L and 0.18 mmol/L respectively (**Table 2.6**) (Teslovich et al., 2010). Furthermore, p.G116S homozygotes do not have the severe LDL-C phenotypes characteristic of homozygous familial hypercholesterolemia in which patients have plasma LDL-C concentrations 6-10 fold normal concentrations (Liyanage et al., 2011). In this context, the effect size associated with p.R730W is relatively agreeable with the observed relationship between variant frequency and effect size (**Figure 2.5**). p.R730W is also of interest particularly as the vast

majority of *LDLR* missense variants have been associated with hypercholesterolemia with only one *LDLR* variant previously linked with LDL-C lowering (Boright et al., 1998). Ultimately, this study has established the association between p.G116S and LDL-C. The effect of either variant on LDLR function as well as the association between the respective effects on LDL-C and CVD-related end-points such as myocardial infarction remains to be determined.

5.1.2 Cardio-metabolic and AD variation in "cognitive impairment, no dementia"

We have described the most comprehensive genetic analysis of the common pre-dementia phenotype known as CIND (Dube et al., 2013). Using the Cardio-Metabochip genotyping array, we utilized a "next-generation" approach to investigate the frequency of genomic variants linked with cardio-metabolic traits in CIND patients and controls (Voight et al., 2012). Although we replicated a potential association between CIND and rs1439568 in the *ZNF608/GRAMD3* locus on chromosome 5 (**Table 3.3**), we did not observe strong evidence linking cardio-metabolic variation in CIND. We also sought to investigate the potential association between established non-ApoE AD-associated variants and CIND. Although no associations were identified, we did replicate similar effect sizes as those previously reported for variants in *CR1*, *ABCA7*, and *PICALM* (**Table 3.4**). Given that our sample size was a small fraction of the large cohorts meta-analyzed in AD GWAS, it is likely that low statistical power hindered our ability to detect small yet significant effect sizes. In order to partly address the issue of limited power, we assessed the accumulation of multiple AD risk alleles in CIND patients compared to controls using a genetic risk score (GRS) utilizing the 11 non-*APOE* AD-

associated variants. Comparable AD-GRSs between CIND patients and controls did not support a role for non-ApoE AD-associated variants as determinants of CIND status. Lastly, we investigated the frequency of the APOE E4 isoform in CIND patients as the E4 isoform remains the strongest genetic determinant of AD. The modest association between the E4 isoform and CIND status compared to controls (OR=1.35, P=0.044, **Table 3.5**) did not provide robust evidence for the E4 isoform as a marker of CIND; however, a previous study identified a 2.7- to 5-fold increase in AD progression from CIND in E4 carriers (Hsiung et al., 2004; Tuokko et al., 2003). In summary, our findings did not support a strong correlation between common cardio-metabolic and AD-associated variants in CIND susceptibility. The phenotypic heterogeneity inherent in CIND, however, is likely to continually confound further genetic analyses unless pre-dementia patients can be better stratified based on the disease underlying the observed cognitive phenotype.

As no previous studies investigated genetic determinants of CIND or related mild cognitive phenotypes with the exception of APOE E4 carrier status, we sought to investigate the role of genetic variation in CIND based on the two most common diseases leading to dementia which remain AD and vascular dementia (VaD). As VaD- and AD-related mechanisms account for the majority of dementia cases among the elderly, we hypothesized that common variants previously associated with cardio-metabolic traits and AD risk were also associated with CIND. The use of the Cardio-Metabochip was essential in testing the former part of this hypothesis by facilitating targeted genotyping of common variation at genome-wide loci associated with VaD-related traits. This type of

trans-disciplinary application represented a novel approach to studying cognitive health and may be applicable to the study of similar diseases affecting cognition. In testing the latter hypothesis that AD-associated variation is also associated with CIND, we developed a panel of 11 AD-associated variants based on the largest AD GWAS meta-analyses (Hollingworth et al., 2011; Naj et al., 2011). The targeted approaches utilized here have represented a multi-disciplinary approach that has not been widely implemented but is likely to gain support as it has become increasingly evident that many cognitive disorders share similar genetic risk loci (Cross-Disorder Group of the Psychiatric Genomics Consortium, 2013). Ultimately, our study provided a preliminary yet comprehensive assessment of genetic determinants of CIND and also highlighted some of the issues involved in studying a pre-clinical cognitive phenotype that must be considered in future analyses of pre-clinical dementia.

5.1.3 Type 2 diabetes-associated common variation in aboriginal populations

We have demonstrated that T2D-associated variants from European and South-Asian populations are also common among aboriginal Canadians and Greenlanders but are not clearly associated with T2D susceptibility. North American aboriginal populations continue to experience significant socio-cultural change and a generally accepted indicator of increasing "Westernization" has been the rapid increase in T2D frequency among aboriginal North Americans. We therefore sought to investigate the frequencies of the top GWAS-identified T2D variants from European and South-Asian cohorts in order to assess whether putative T2D loci are also applicable within aboriginal populations (**Table 4.3**). Our test for association between T2D variants and T2D status within the

Sandy Lake Oji-Cree and Inuvik Inuit did not reveal strong correlations, however, this analysis was limited by both the sample size and the small effect sizes attributed to each variant regarding T2D risk (**Table 4.5**). We also tested for association between the T2D variants and variability in fasting blood glucose (FBG) within three populations of Inuit descent which revealed a modest association between the rs7178572 variant near *HMG20A* on chromosome 15 in a combined analysis (**Table 4.6**). We further assessed the accumulation of multiple T2D variants through a T2D GRS (**Table 4.7**). While mean GRS based on all 17 selected variants or only European-identified variants were similar between T2D patients and controls, we observed a nominally significant difference in mean GRS between Inuvik T2D patients and controls based on South Asian-associated variants. Cumulatively, these analyses did not provide a strong indication that established T2D variation is associated with T2D among aboriginals. However, we have provided a first look into the frequency and effect sizes of European- and South-Asian-identified T2D-associated variation as it applies to aboriginal Canadians and Greenlanders.

Similar to our approach for studying CIND genetics, we utilized large-scale GWAS meta-analyses to identify a panel of the top T2D-associated variants. This approach allowed us to focus our investigation on established T2D loci which was the most feasible means for assessing genetic T2D risk amongst aboriginal populations given limited sample sizes. Thousands of T2D patients and controls from homogenous sample populations were required in order to identify genome-wide significant associations between common variants and T2D as well as T2D-related phenotypes. As no analogous aboriginal cohorts exist that would adequately power the discovery of additional T2D-

associated variants, our alternate approach took advantage of the wealth of T2D GWAS data to propose candidate loci. To further address the issue of limited sample size, we combined aboriginal populations of common descent when investigating FBG among the Inuit. A similar approach may be required in future genetic studies within aboriginal populations in order to improve sample size and statistical power.

An additional consideration for our study related to the investigation of common variation in multi-ethnic cohorts. A fundamental requirement for GWAS involves the use of homogenous study populations which is usually addressed in part by stratification based on ethnicity. Replication of GWAS results in additional multi-ethnic cohorts represents a key step in confirming potential gene-disease associations. In addition to our study limitations, further ethnicity-specific factors have been hypothesized to affect replicability in multi-ethnic association studies using common variants. These factors may include gene-gene and gene-environment interactions but ethnicity-specific patterns of linkage disequilibrium are believed to largely account for the discrepancies observed in multi-ethnic association studies (Fu et al., 2011; Ioannidis et al., 2004; Lanktree et al., 2009; Lin et al., 2007). While the more pertinent limitations of our study involved issues of sample size, future genetic studies aimed at replicating associations in aboriginal populations must consider the role of ethnicity-specific patterns of linkage disequilibrium as a potential confounding factor.

5.2 Current methodological limitations

The most significant limitation hindering genetic studies of complex disease relates fundamentally to the uncertainty surrounding the concept of complex disease genetics. Based on GWAS results from recent years, the portion of complex disease heritability explained by common variation has fallen short of what was initially expected as outlined in the "common disease-common variant" (CDCV) hypothesis (Reich and Lander, 2001). Furthermore, the tendency of GWAS to identify variants localized in gene deserts and intronic or intergenic regions has complicated the translation of robust association signals to biological relevance or clinical utility. Accordingly, new hypotheses have recently emerged that aim to account for the "missing heritability" in complex disease susceptibility that GWAS has failed to uncover. These hypotheses have largely supported the pooling of GWAS data for large-scale meta-analyses, a shift towards investigating the role of rare variation in complex disease susceptibility, and further *in vitro* and *in vivo* modeling of variants.

5.2.1 The CDCV hypothesis then and now

As previously discussed, the CDCV hypothesis provided the first generally accepted concept for the role of common variation in common, complex disease susceptibility. As the CDCV model began to formulate in the late 1990's, it was believed that the anticipated catalogue of common human genomic variants would be used to perform hundreds of thousands of association tests in what became known as GWAS. A decade since the initial draft human genome release, GWAS have been applied to virtually every

common complex disease and the overall results were not quite as definitive as expected. As GWAS have unanimously revealed, associated variants individually contribute modestly to disease risk across almost all complex diseases. Combined analyses involving the top GWAS hits from multiple loci for a disease, in the form of risk scores, still only explain a modest percentage of disease heritability. The portion of complex disease heritability left unexplained by common variation has been dubbed the "missing heritability" (Manolio et al., 2009). Based on the small effect sizes assigned to associated variants as well as the stringent Bonferroni-corrected significance thresholds required when performing $>10^6$ tests for association, GWAS require thousands of carefully phenotyped cases and controls in order to support the likelihood of detecting association. Despite validating candidate disease loci as well as identifying many novel and unexpected risk loci, the hunt for the "missing heritability" has come to dominate the continued effort to understand complex disease etiology.

Going forward, the lessons learned from GWAS must be utilized to re-assess and improve the current working model of complex disease genetics. Due largely to the issue of missing heritability, the CDCV hypothesis has recently been revisited and refined. Two main approaches have been described which focus on the continued search for common variant associations and on the role of rare variation in complex disease susceptibility. The potential remains for additional undiscovered common variants of subtle effect to contribute to disease susceptibility but it is believed that GWAS have simply been underpowered to detect these associations. Targeted genotyping and resequencing efforts have been proposed to help reveal additional common variants while

lowering the stringent statistical requirements typical of GWAS by focusing on select genomic loci. Similarly, meta-analyses of GWAS data are also likely to reveal previously undetected associations due to increased sample sizes and statistical power. Alternatively, it is also hypothesized that GWAS-identified variants are tagging low-frequency variants – the type that are excluded from GWAS – through linkage disequilibrium. These rare variants, of minor allele frequency <1%, are thought to have greater effects on disease susceptibility and may contribute to the "missing heritability" puzzle. While these approaches aim to account for the "missing heritability", it has also been suggested that the total heritability ascribed to a disease may be overestimated. By failing to account for gene-gene or epistatic interactions, total heritability estimates may actually be inflated and thus create "phantom heritability" that cannot be explained by the discovery of additional variant associations (Zuk et al., 2012). A general road map for the future of genomics in complex disease investigation has thus been proposed, however, the extent to which these approaches will address current limitations remains to be observed.

5.2.2 Clinical translation of GWAS findings

A commonly discussed limitation involves the lack of clinical utility in GWAS results. As GWAS-identified variants currently explain a small proportion of disease risk or phenotypic variability, it has been difficult utilizing GWAS findings to help identify high-risk individuals or to re-classify at-risk patients. Although this is a significant limitation for the immediate application of GWAS results, the true worth seems to lie in the ability for GWAS to nominate disease-associated loci. Identifying the genes involved in disease pathogenesis may prove to be just as important as the discovery of a high-risk

variant as detailed understanding into disease mechanisms and the biological players involved will facilitate the development of therapeutic strategies. To appreciate this concept, one may look no further than the example of the targeted inhibition of 3-hydroxy-3-methyl-glutaryl-CoA reductase (HMGCR) by statin therapy. As HMGCR is the rate-limiting enzyme in endogenous cholesterol synthesis, its inhibition markedly lowers plasma LDL-C and significantly lowers CAD risk (Brugts et al., 2009). Common variation at the *HMGCR* locus has been associated with a modest effect on plasma cholesterol yet inhibition of this key enzyme has shown profound effects on cholesterol homeostasis which have translated to cardiovascular benefits. Similarly, the GWAS-identified sortilin gene, encoded by *SORT1*, has been revealed as a novel receptor involved in LDL-C homeostasis and suggests a potential therapeutic target. As many common complex diseases have been associated with multiple loci through GWAS meta-analyses, the monumental task of investigating the explanation underlying these association signals is currently underway. Thus the clinical value of GWAS may be forthcoming as disease etiology is better understood.

5.2.3 An end to the GWAS era?

With many common complex diseases now investigated by the >1400 published GWAS, the need for the execution of additional GWAS has been questioned largely on the grounds that 1) multi-ethnic GWAS data now exist for many common complex diseases such as CVD and AD; 2) existing GWAS datasets must be investigated for biological relevance; and 3) next-generation genotyping platforms have been designed for enriched genotyping of variants in candidate loci or exonic regions based on GWAS meta-analyses

which will permit more direct hypothesis testing. Much remains to be investigated through association studies; however, the traditional GWAS approach must be adapted to the evolving concept of genetic susceptibility in complex disease etiology in a manner on pace with the continued development of the CDCV hypothesis.

5.3 Future directions for genomic analyses of complex disease

As better characterization of the current unexplained heritability will be investigated through assessment of rare variation, a variety of strategies – both novel and established techniques – are likely to become increasingly important to future genomic studies of complex disease. The increasing feasibility of next-generation sequencing (NGS) platforms has made whole genome and exome sequencing more accessible where NGS is expected to become the new standard approach in variant discovery. Established approaches such as studies using population isolates, monogenic disease phenotypes and investigations into the extremes of quantitative trait distributions also remain viable strategies in the discovery of susceptibility loci. More experimental approaches may also prove insightful such as the use of pre-clinical mouse models to nominate candidate loci in human diseases. Together, this range of techniques promises to help better characterize the genetic architecture underlying complex disease susceptibility.

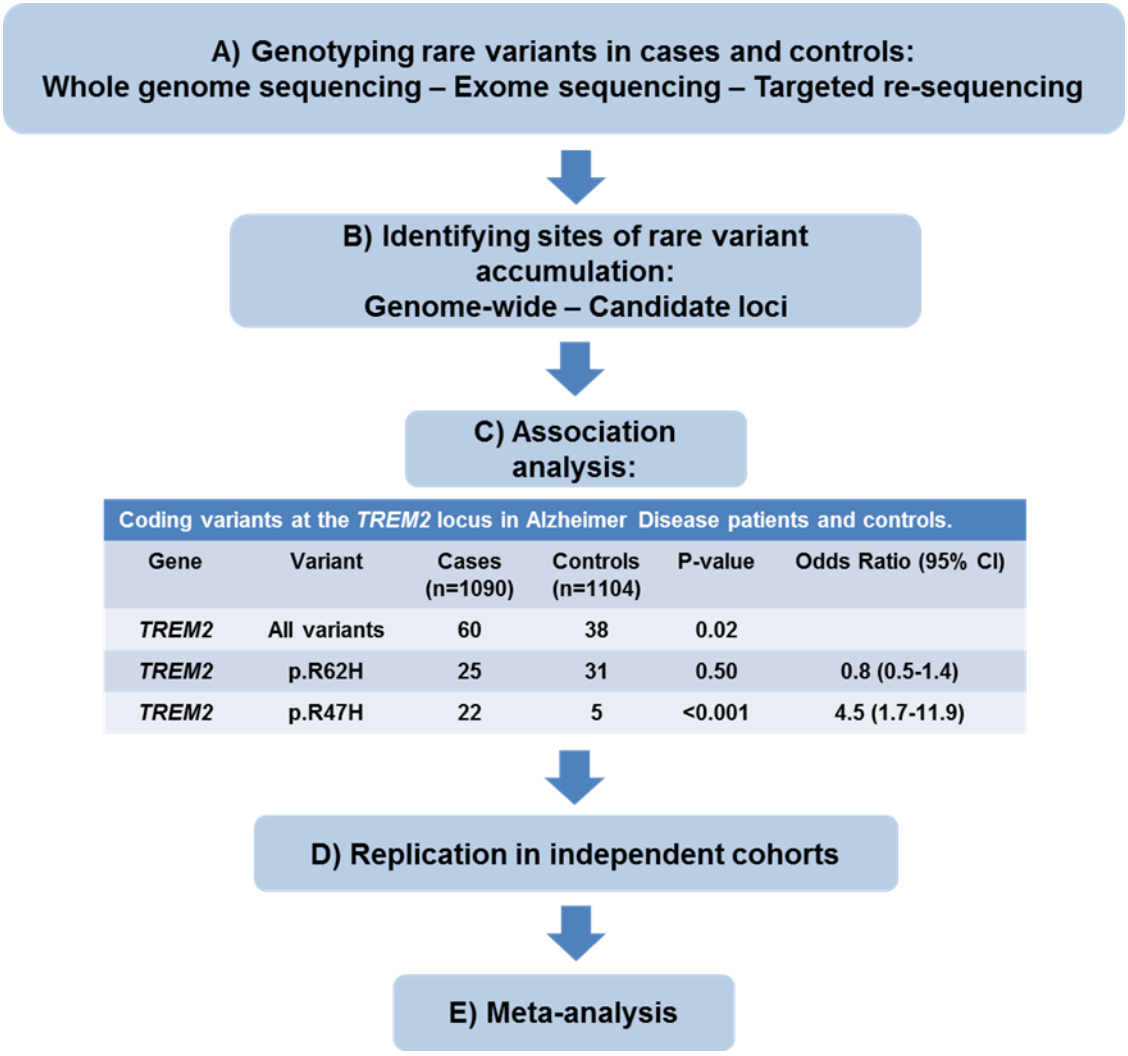
5.3.1 Next-generation sequencing

As only ~10% of common variation is assessed using GWAS panels through the use of SNPs (Willer and Mohlke, 2012), whole genome and exome sequencing have become

increasingly implemented for the utility offered by the comprehensive assessment of genomic variation. Due to the costs of NGS, current applications have been limited in sample size relative to GWAS. As a result, a two-stage study design has emerged as a common workflow. First, small discovery cohorts of cases and controls are sequenced and all genetic variants are identified. The subsequent partitioning of variants based on frequency or genomic position can be used to nominate candidate variants for genotyping in a larger replication cohort (Kang et al., 2012). Alternatively, exome sequencing offers a focused approach to sequencing only protein-coding regions of the genome where variation may be more likely to have deleterious effects.

The recent discovery of a novel large-effect AD-associated variant in the *TREM2* gene independently by two groups illustrates the potential for NGS-based association studies. Using NGS in a case-control design, Guerreiro *et al.* were able to target loci harbouring a significant accumulation of rare variants which were then replicated in publicly available GWAS datasets (Guerreiro et al., 2013). Jonsson *et al*. independently identified the novel AD-associated variant in *TREM2*; however, this was accomplished by first comprehensively characterizing genetic variation in a genetically homogeneous Icelandic population (Jonsson et al., 2013). As described earlier, the use of population isolates in association studies represents an established approach which limits genetic heterogeneity and improves statistical power (Tian et al., 2008). These workflows are likely to represent a recurring strategy applied in future NGS-based studies of complex diseases and are summarized in **Figure 5.1**.

Figure 5.1 Investigating rare variation in complex disease. Using recent rare variation studies in AD as a template (Guerreiro et al., 2013; Jonsson et al., 2013), a common workflow is emerging for studying the role of rare variation in complex disease. **A)** Next-generation sequencing (NGS) of the whole genome or the exome is used to comprehensively identify genomic variation. This is performed in a small number of participants or cases and controls as a variant discovery phase. **B)** Variation can be investigated agnostically by scanning the entire genome or using a hypothesis-driven approach where candidate loci may be prioritized as sites believed to be harbouring a burden of variation. **C)** Rare variants are selected based on allele frequency (MAF<1%). Association analyses begin with a gene-based approach where the accumulation of rare variants at a given gene or locus is compared between cases and controls. In this example, AD cases had a significantly higher accumulation of rare variants in *TREM2* versus controls. Closer investigation revealed that the R47H variant was associated with AD status. **D)** Using publicly available AD GWAS datasets, genotyping data on millions of common genomic variants can be used to replicate associations discovered using NGS through the direct genotyping of the variant of interest or through imputation. **E)** A final estimate of the variant effect size and frequency can be determined by combining all datasets into a single statistical analysis or meta-analysis.



Coding variants at the *TREM2* locus in Alzheimer Disease patients and controls.

Gene	Variant	Cases (n=1090)	Controls (n=1104)	P-value	Odds Ratio (95% CI)
TREM2	All variants	60	38	0.02	
TREM2	p.R62H	25	31	0.50	0.8 (0.5-1.4)
TREM2	p.R47H	22	5	<0.001	4.5 (1.7-11.9)

On the clinical side, NGS has proven useful in helping to diagnose patients with suspected genetic disorders when candidate re-sequencing approaches have failed. For example, Rios *et al.* used whole genome sequencing to correctly identify the disease-causing mutations underlying an 11-month-old girl's sitosterolemia after being misdirected by an initial presentation of hypercholesterolemia (Rios et al., 2010). In an analogous case, exome sequencing helped elucidate the causal variant underlying AD in a patient from a consanguineous family when candidate sequencing of known AD-associated genes including *APP, PSEN1* and *PSEN2*, failed to identify potential disease-causing variants (Guerreiro et al., 2012). The subsequent exome sequencing ultimately identified a variant in *NOTCH3* which was previously associated with cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy (CADASIL); the most common hereditary form of stroke (Joutel et al., 1996). Instances of NGS approaches applied to solving unusual clinical cases have become increasingly reported as the comprehensive nature of genome-wide sequencing provides clinicians with unprecedented insight into genetic disease pathogenesis and ultimately improved diagnostic capability.

5.3.2 Lessons from monogenic diseases and extreme phenotypes

Linkage studies of monogenic disorders established the sufficiency of a single deleterious variant to cause remarkable phenotypes such as *LDLR* mutations in FH patients. By elucidating the susceptibility loci in monogenic diseases, we gain a fundamental insight into the biological pathways involved in modulating disease susceptibility and phenotypic variability that can be applied to patients with less severe but unhealthy phenotypes.

Combining the study of monogenic disorders with emerging genomic technologies, it will be possible to uncover candidate loci that have eluded GWAS but nonetheless may have a profound phenotypic effect if perturbed. Furthermore, the process of identifying disease variants has become streamlined with the usage of NGS platforms as investigation is no longer limited to candidate loci but now incorporates the majority of genomic variation (Gilissen et al., 2012). Another emerging approach involves the dichotomizing of quantitative traits by sampling disease-free participants that occupy the tail-ends, or extremes, of trait distributions. While GWAS have been applied to the extremes of phenotypic distributions, NGS approaches have not been widely utilized in this respect and may offer new insight into the genetic architecture and biological pathways involved in modulating phenotypic variability (Barnett et al., 2012).

5.3.3 Mouse disease models and candidate susceptibility loci

Model organisms, particularly murine models, play a key role in the development of therapeutic strategies and are often utilized in identifying the effects of treatment in complex disease models (Welch, 2012). Alternatively, murine models can also be used to identify novel loci in complex human diseases. Studies using congenic mapping approaches in mice expressing a human disease phenotype such as atherosclerosis or hypertriglyceridemia have successfully implicated mouse susceptibility loci in human disease. Gargalovic *et al.* identified an association between variation in the *Zhx2* gene with plasma lipid metabolism using a congenic mapping technique; a locus that had not previously been implicated in human lipid metabolism (Gargalovic et al., 2010). Interestingly, a subsequent GWAS meta-analysis on carotid intima media thickness (C-

IMT), a measure of subclinical atherosclerosis, revealed an association with *ZHX2* which proposed the potential use of mouse genetics to better understand the genetic architecture of complex phenotypes in humans (Bis et al., 2011). Using a different approach, Johansen *et al*. showed that patients with polygenic hypertriglyceridemia (HTG) carried a significant burden of rare variants at *GPIHBP1* and *LMF1* which were first associated with murine HTG phenotypes (Johansen et al., 2012). Thus the application of mouse genetics in identifying novel susceptibility loci in humans may prove effective in the investigation of additional complex phenotypes.

5.4 Personalized medicine and therapeutic strategies

In brief, the concept of personalized medicine has centered on the design and implementation of health care tailored to the individual patient's unique biological and genetic components. With the increasing feasibility of whole genome sequencing, there has been growing anticipation surrounding the idea that genomic data will facilitate improved patient care in the near future. The incomplete understanding between genotype and complex disease risk, however, remains the greatest limiting factor in the integration of genomic data for the improvement of personalized medicine. Despite this limitation, GWAS and emerging genomic techniques have impacted upon the current concept of personalized medicine and have shed new light on future genomic applications in personalized medicine. GWAS have undeniably helped to advance the concept of personalized medicine as well as the field of pharmacogenomics and drug design which has suggested that the legacy of GWAS has yet to be fully realized.

5.4.1 Personalized medicine in the genomics era

The formation of risk scores based on GWAS findings served as an initial attempt at validating a role for common variants in assessing patient risk. Aggregate scores of risk-associated alleles from the top GWAS-identified variants associated with a specific phenotype were calculated in order to assess the genetic risk burden within the patient. These risk scores were then tested in independent cohorts of patients with related phenotypes in order to validate the clinical utility of such risk scores with the intention of incorporating genetic information within traditional risk algorithms such as the Framingham Heart Score. Modest success has been attributed to genetic risk scores that assess complex disease risk including CVD and AD (Kathiresan et al., 2008; Rodriguez-Rodriguez et al., 2012). In one relatively successful study, Kathiresan *et al*. showed that a GRS of 9 variants associated with LDL or HDL cholesterol was an independent risk factor for incident CVD, however, genotype score did not improve clinical risk prediction (Kathiresan et al., 2008). Associations between GRS and a given phenotype have been commonly reported, however the limited ability for GRS to substantially reclassify at-risk patients has largely dissuaded the clinical utility of genetic risk prediction in the general population (Jostins and Barrett, 2011). Due to the small effects on risk that are ascribed to GWAS-identified loci, it may not be surprising that panels of small-effect variants do not significantly improve patient risk prediction. With genomic studies shifting from common variation toward rare variation, the identification of rare variants with potentially larger effects on risk may be more suited for clinical utility in calculating complex disease risk in the patient.

As our understanding of the genetic determinants of complex disease progresses, it will become increasingly pertinent to assess the patient's predisposition to disease at the earliest time point which is commonly perceived to be *in utero*. Prenatal genomic analysis represents a rapidly developing field within personalized medicine where the assessment of disease susceptibility may begin during fetal development (Bianchi, 2012). Until recently, prenatal diagnostic techniques were limited to ultrasonography and fetal metaphase karyotyping. The more sophisticated analyses used today have incorporated DNA micro-array-based assays implemented in conjunction with karyotyping to identify high-risk chromosomal abnormalities. Historically, however, fetal health has largely been based on morphological factors and low-resolution genetic analyses. The recent discovery that the fetal genome can be sequenced non-invasively from maternal blood has opened the door for prenatal screening techniques that utilize current and developing genomic technology (Fan et al., 2012). Sequencing of the fetal genome offers variant detection at the highest resolution and thus provides the means for identifying potentially deleterious point mutations. The utility of this information in personalized medicine theoretically has great potential. The effectiveness of this strategy is ultimately dependent upon our ability to interpret variation as being deleterious or benign. Thus the quick succession of technological advances in genetics has made it easy for our reach to exceed our grasp in terms of the ability to generate data but the limitation in using the data for clinical decision-making.

5.4.2 Defining the "genomics" in pharmacogenomics

Heterogeneity in patient response to drug treatment remains a major complication in delivering cost-effective health care. While trial-and-error may be used for some drugs in order to identify the optimal dosage for the patient, this process is imprecise, costly, ineffective if the patient cannot metabolize the drug, and potentially hazardous. The genetic component underlying the variability in drug response has long been suspected, however the genomic techniques described here have only recently been established in studies on heterogeneity in drug response as well as susceptibility to adverse reactions. As with the study of complex phenotypes, candidate gene studies provided the first albeit limited investigations into genetic determinants of variable drug response from which the term pharmacogenetics was coined. GWAS have helped expand the concept of the genetic architecture underlying drug response while next-generation approaches such as whole-genome and exome sequencing are poised to provide even greater detail regarding the genetic components involved in modulating pharmacologic effects.

Statin therapy (3-hydroxy-3methylglutaryl-coenzyme A reductase [HMGCR] inhibitors) represents a prime example of a drug treatment that has become better understood through modern genomic approaches. Despite their status as the standard drug treatment for lowering LDL-C and thus cardiovascular risk, statin-mediated LDL-C lowering can vary as much as 10% to 70% in the case of rosuvastatin (Simon et al., 2006). Accordingly, an important future goal for the prescription of statin drugs has focused on the identification of genetic factors that may help in determining the optimal statin and drug dosage tailored to the patient. Candidate gene studies importantly

identified a common haplotype in *HMGCR* encoding an alternatively spliced gene product associated with reduced LDL-C response to simvastatin (Krauss et al., 2008). Subsequent candidate gene studies and GWAS implicated several additional loci which helped characterize pathways relevant to statin pharmacodynamics (Chasman et al., 2012; Voora and Ginsburg, 2012). Although association signals have been reported at *PCSK9*, *ABCG2*, *LPA*, the most consistent findings have associated APOE isoforms with variable statin-mediated LDL-C response; carriers of the E2 isoform are associated with the greatest LDL-C lowering followed by E3 and E4 isoforms (Voora and Ginsburg, 2012). Although these findings have not yet translated to changes in the process of statin prescription, GWAS have importantly identified some of the biological players and pathways involved in statin uptake and efficacy which will inform future drug design.

Genomic approaches may also benefit studies on adverse drug reactions (ADRs). Again, GWAS on statin myopathy, or muscle pain and weakness due to statin therapy (Thompson et al., 2003), successfully identified a robust association between common variation at the *SLCO1B1* gene and statin myopathy susceptibility (OR=16.9; 95% CI=4.7-61.1) (Link et al., 2008). Despite suggestions from the Food and Drug Administration as well as the Clinical Pharmacogenetics Implementation Consortium to institute clinical genotyping of *SLCO1B1* variants to assess patient risk, statin myopathy risk continues to be largely managed using trial-and-error and monitoring of serum creatine kinase levels (Wilke et al., 2012). Further examples of the potential benefit of pharmacogenomics in ADR studies abound (Harper and Topol, 2012). For instance, prospective screening of the HLA haplotype HLA-B*5701 in patients taking abacavir, an

inhibitor of human immunodeficiency virus (HIV) reverse transcriptase, was shown to reduce the frequency of abacavir-related hypersensitivity reactions from 3% to 0% (Mallal et al., 2008). As studies continue to make a strong case for the integration of pharmacogenomics in patient care, it seems increasingly plausible that genomic data will have a greater role in determining safe and effective drug dosage.

The next stage for pharmacogenomics is likely to involve greater utilization of exome or whole genome sequencing for the identification of rare variants that cannot be tested for association using traditional GWAS methods. As pharmacogenomics studies using NGS have yet to be published, we can only speculate on the potential for NGS to advance pharmacogenomics. However, large-scale resequencing studies have provided some insight into what may be expected from next-generation studies. One resequencing study by Nelson *et al.* reported an abundance of rare variation in 202 genes encoding drug targets where rare variants were observed in ~1 in every 17 bases (Nelson et al., 2012). Additionally, Ramirez *et al*. reported an abundance of rare variants in patients with drug-induced long QT syndrome (diLQTS) at loci associated with congenital arrhythmia syndrome suggesting that rare variation at known arrhythmia loci plays a role in diLQTS predisposition (Ramirez et al., 2012). Analogous results may be expected when next-generation methods are further integrated into pharmacogenomics, however the transition from variant discovery to clinical incorporation remains contested.

5.4.3 Pharmacological design

GWAS and resequencing studies may also prove to be powerful tools in the complicated process of drug design. As discussed, GWAS have provided valuable insights into the genetic component underlying many common and complex diseases thus improving our concept of the biological pathways implicated in any given complex disease. By modulating gene expression at GWAS-identified susceptibility loci, it may be possible to produce a potentially therapeutic effect. Currently, the technologies exist whereby small molecule inhibitors, anti-sense oligonucleotides (ASOs) and gene replacement can effectively target and perturb gene expression and have already been incorporated in emerging therapeutic strategies.

The leading-edge of dyslipidemia therapies is represented by a host of novel treatment strategies that each utilizes pharmacologic technologies. Lomitapide, a small molecule inhibitor of microsomal triglyceride transfer protein (MTTP), was developed as a cholesterol-lowering therapy that recently received FDA approval for the treatment of homozygous familial hypercholesterolemia (Cuchel et al., 2007). Mipomersen, another cholesterol-lowering therapy, is an ASO designed to hybridize to and degrade apoB mRNA thus reducing expression of apoB expression and, subsequently LDL cholesterol (Ricotta and Frishman, 2012). Advances in viral gene transfer technologies have led to breakthroughs in gene therapy research. Alipogene tiparvovec represents the first gene therapy approved for marketing in Europe for the treatment of familial lipoprotein lipase deficiency (LPLD) (Dube and Hegele, 2012). LPLD patients lack a fully functional copy of lipoprotein lipase (LPL) thus alipogene tiparvovec partially restores functional LPL

through transient transduction by viral particles containing a functional copy of human *LPL* (Dube and Hegele, 2012). The emerging gene therapies have been well-suited to treating monogenic disorders. In translating gene therapy to complex polygenic diseases, however, the difficulty remains with identifying the ideal genomic targets which will require deciphering the biological relevance of the top GWAS genes.

5.5 Conclusions

Genetic studies of complex human disease are set to undergo a dramatic shift in both the technological and analytical approaches used to evaluate genetic risk. The studies described here encapsulate the current techniques that have been utilized to develop and test our current understanding of genetic susceptibility to common and complex disease. We have demonstrated 1) a candidate gene resequencing study in which the private common *LDLR* variant G116S was associated with plasma LDL-C among Inuit descendants; 2) the design and execution of a targeted GWAS investigating the role of cardio-metabolic and AD-associated variation in pre-dementia susceptibility; and 3) a candidate genotyping study of T2D-associated GWAS variants in aboriginal Canadian and Greenlander populations. Collectively, these three studies represent the established techniques implemented in assessing the genetic architecture underlying complex phenotypes. Utilizing these techniques, we have contributed new insight into the genetic component underlying plasma LDL-C concentration and cognitive decline as well as the frequency of T2D risk alleles in aboriginal Canadians. With the emergence of NGS, rare variant analysis has come to represent the shift in focus with the aim of accounting for

some of the disease heritability left unexplained by GWAS. The next chapter in the genomic study of human disease will undoubtedly require a new set of analytical procedures, however the lessons learned from both classical genetics and the recent CDCV hypothesis-driven era have been invaluable in establishing a concept of genetic risk in complex disease upon which we can continue to build.

5.6 References

Barnett, I.J., Lee, S., and Lin, X. (2012). Detecting Rare Variant Effects Using Extreme Phenotype Sampling in Sequencing Association Studies. *Genet Epidemiol*.

Bianchi, D.W. (2012). From prenatal genomic diagnosis to fetal personalized medicine: progress and challenges. *Nat Med 18*, 1041-1051.

Bis, J.C., Kavousi, M., Franceschini, N., Isaacs, A., Abecasis, G.R., Schminke, U., Post, W.S., Smith, A.V., Cupples, L.A., Markus, H.S., *et al.* (2011). Meta-analysis of genome-wide association studies from the CHARGE consortium identifies common variants associated with carotid intima media thickness and plaque. *Nat Genet 43*, 940-947.

Boright, A.P., Connelly, P.W., Brunt, J.H., Morgan, K., and Hegele, R.A. (1998). Association and linkage of LDLR gene variation with variation in plasma low density lipoprotein cholesterol. *J Hum Genet 43*, 153-159.

Brugts, J.J., Yetgin, T., Hoeks, S.E., Gotto, A.M., Shepherd, J., Westendorp, R.G., de Craen, A.J., Knopp, R.H., Nakamura, H., Ridker, P., *et al.* (2009). The benefits of statins in people without established cardiovascular disease but with cardiovascular risk factors: meta-analysis of randomised controlled trials. *BMJ 338*, b2376.

Chasman, D.I., Giulianini, F., MacFadyen, J., Barratt, B.J., Nyberg, F., and Ridker, P.M. (2012). Genetic determinants of statin-induced low-density lipoprotein cholesterol reduction: the Justification for the Use of Statins in Prevention: an Intervention Trial Evaluating Rosuvastatin (JUPITER) trial. *Circ Cardiovasc Genet 5*, 257-264.

Cross-Disorder Group of the Psychiatric Genomics Consortium, T. (2013). Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet*.

Cuchel, M., Bloedon, L.T., Szapary, P.O., Kolansky, D.M., Wolfe, M.L., Sarkis, A., Millar, J.S., Ikewaki, K., Siegelman, E.S., Gregg, R.E., *et al.* (2007). Inhibition of microsomal triglyceride transfer protein in familial hypercholesterolemia. *N Engl J Med 356*, 148-156.

Dube, J.B., and Hegele, R.A. (2012). The application of gene therapy in lipid disorders: where are we now? *Clin Lipidol 7*, 419-429.

Dube, J.B., Johansen, C.T., Robinson, J.F., Lindsay, J., Hachinski, V., and Hegele, R.A. (2013). Genetic determinants of "cognitive impairment, no dementia". *J Alzheimers Dis 33*, 831-840.

Fan, H.C., Gu, W., Wang, J., Blumenfeld, Y.J., El-Sayed, Y.Y., and Quake, S.R. (2012). Non-invasive prenatal measurement of the fetal genome. *Nature 487*, 320-324.

Fu, J., Festen, E.A., and Wijmenga, C. (2011). Multi-ethnic studies in complex traits. *Hum Mol Genet 20*, R206-213.

Gargalovic, P.S., Erbilgin, A., Kohannim, O., Pagnon, J., Wang, X., Castellani, L., LeBoeuf, R., Peterson, M.L., Spear, B.T., and Lusis, A.J. (2010). Quantitative trait locus mapping and identification of Zhx2 as a novel regulator of plasma lipid metabolism. *Circ Cardiovasc Genet 3*, 60-67.

Gilissen, C., Hoischen, A., Brunner, H.G., and Veltman, J.A. (2012). Disease gene identification strategies for exome sequencing. *Eur J Hum Genet 20*, 490-497.

Guerreiro, R., Wojtas, A., Bras, J., Carrasquillo, M., Rogaeva, E., Majounie, E., Cruchaga, C., Sassi, C., Kauwe, J.S., Younkin, S., *et al.* (2013). TREM2 variants in Alzheimer's disease. *N Engl J Med 368*, 117-127.

Guerreiro, R.J., Lohmann, E., Kinsella, E., Bras, J.M., Luu, N., Gurunlian, N., Dursun, B., Bilgic, B., Santana, I., Hanagasi, H., *et al.* (2012). Exome sequencing reveals an unexpected genetic cause of disease: NOTCH3 mutation in a Turkish family with Alzheimer's disease. *Neurobiol Aging 33*, 1008 e1017-1023.

Harper, A.R., and Topol, E.J. (2012). Pharmacogenomics in clinical practice and drug development. *Nat Biotechnol 30*, 1249.

Hollingworth, P., Harold, D., Sims, R., Gerrish, A., Lambert, J.C., Carrasquillo, M.M., Abraham, R., Hamshere, M.L., Pahwa, J.S., Moskvina, V., *et al.* (2011). Common variants at ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP are associated with Alzheimer's disease. *Nat Genet 43*, 429-435.

Hsiung, G.Y., Sadovnick, A.D., and Feldman, H. (2004). Apolipoprotein E epsilon4 genotype as a risk factor for cognitive decline and dementia: data from the Canadian Study of Health and Aging. *CMAJ 171*, 863-867.

International Hapmap Consortium, T. (2003). The International HapMap Project. *Nature 426*, 789-796.

Ioannidis, J.P., Ntzani, E.E., and Trikalinos, T.A. (2004). 'Racial' differences in genetic effects for complex diseases. *Nat Genet 36*, 1312-1318.

Johansen, C.T., Wang, J., McIntyre, A.D., Martins, R.A., Ban, M.R., Lanktree, M.B., Huff, M.W., Peterfy, M., Mehrabian, M., Lusis, A.J., *et al.* (2012). Excess of rare variants in non-genome-wide association study candidate genes in patients with hypertriglyceridemia. *Circ Cardiovasc Genet 5*, 66-72.

Jonsson, T., Stefansson, H., Steinberg, S., Jonsdottir, I., Jonsson, P.V., Snaedal, J., Bjornsson, S., Huttenlocher, J., Levey, A.I., Lah, J.J., *et al.* (2013). Variant of TREM2 associated with the risk of Alzheimer's disease. *N Engl J Med 368*, 107-116.

Jostins, L., and Barrett, J.C. (2011). Genetic risk prediction in complex disease. *Hum Mol Genet 20*, R182-188.

Joutel, A., Corpechot, C., Ducros, A., Vahedi, K., Chabriat, H., Mouton, P., Alamowitch, S., Domenga, V., Cecillion, M., Marechal, E., *et al.* (1996). Notch3 mutations in CADASIL, a hereditary adult-onset condition causing stroke and dementia. *Nature 383*, 707-710.

Kang, G., Lin, D., Hakonarson, H., and Chen, J. (2012). Two-stage extreme phenotype sequencing design for discovering and testing common and rare genetic variants: efficiency and power. *Hum Hered 73*, 139-147.

Kathiresan, S., Melander, O., Anevski, D., Guiducci, C., Burtt, N.P., Roos, C., Hirschhorn, J.N., Berglund, G., Hedblad, B., Groop, L., *et al.* (2008). Polymorphisms associated with cholesterol and risk of cardiovascular events. *N Engl J Med 358*, 1240-1249.

Krauss, R.M., Mangravite, L.M., Smith, J.D., Medina, M.W., Wang, D., Guo, X., Rieder, M.J., Simon, J.A., Hulley, S.B., Waters, D., *et al.* (2008). Variation in the 3-hydroxyl-3-methylglutaryl coenzyme a reductase gene is associated with racial differences in low-density lipoprotein cholesterol response to simvastatin treatment. *Circulation 117*, 1537-1544.

Lanktree, M.B., Anand, S.S., Yusuf, S., and Hegele, R.A. (2009). Replication of genetic associations with plasma lipoprotein traits in a multiethnic sample. *J Lipid Res 50*, 1487-1496.

Lin, P.I., Vance, J.M., Pericak-Vance, M.A., and Martin, E.R. (2007). No gene is an island: the flip-flop phenomenon. *Am J Hum Genet 80*, 531-538.

Link, E., Parish, S., Armitage, J., Bowman, L., Heath, S., Matsuda, F., Gut, I., Lathrop, M., and Collins, R. (2008). SLCO1B1 variants and statin-induced myopathy--a genomewide study. *N Engl J Med 359*, 789-799.

Liyanage, K.E., Burnett, J.R., Hooper, A.J., and van Bockxmeer, F.M. (2011). Familial hypercholesterolemia: epidemiology, Neolithic origins and modern geographic distribution. *Crit Rev Clin Lab Sci 48*, 1-18.

Mallal, S., Phillips, E., Carosi, G., Molina, J.M., Workman, C., Tomazic, J., Jagel-Guedes, E., Rugina, S., Kozyrev, O., Cid, J.F., *et al.* (2008). HLA-B*5701 screening for hypersensitivity to abacavir. *N Engl J Med 358*, 568-579.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., *et al.* (2009). Finding the missing heritability of complex diseases. *Nature 461*, 747-753.

Naj, A.C., Jun, G., Beecham, G.W., Wang, L.S., Vardarajan, B.N., Buros, J., Gallins, P.J., Buxbaum, J.D., Jarvik, G.P., Crane, P.K., *et al.* (2011). Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease. *Nat Genet 43*, 436-441.

Nelson, M.R., Wegmann, D., Ehm, M.G., Kessner, D., St Jean, P., Verzilli, C., Shen, J., Tang, Z., Bacanu, S.A., Fraser, D., *et al.* (2012). An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science 337*, 100-104.

Ramirez, A.H., Shaffer, C.M., Delaney, J.T., Sexton, D.P., Levy, S.E., Rieder, M.J., Nickerson, D.A., George, A.L., Jr., and Roden, D.M. (2012). Novel rare variants in congenital cardiac arrhythmia genes are frequent in drug-induced torsades de pointes. *Pharmacogenomics J*.

Reich, D.E., and Lander, E.S. (2001). On the allelic spectrum of human disease. *Trends Genet 17*, 502-510.

Ricotta, D.N., and Frishman, W. (2012). Mipomersen: a safe and effective antisense therapy adjunct to statins in patients with hypercholesterolemia. *Cardiol Rev 20*, 90-95.

Rios, J., Stein, E., Shendure, J., Hobbs, H.H., and Cohen, J.C. (2010). Identification by whole-genome resequencing of gene defect responsible for severe hypercholesterolemia. *Hum Mol Genet 19*, 4313-4318.

Rodriguez-Rodriguez, E., Sanchez-Juan, P., Vazquez-Higuera, J.L., Mateo, I., Pozueta, A., Berciano, J., Cervantes, S., Alcolea, D., Martinez-Lage, P., Clarimon, J., *et al.* (2012). Genetic risk score predicting accelerated progression from mild cognitive impairment to Alzheimer's disease. *J Neural Transm*.

Simon, J.A., Lin, F., Hulley, S.B., Blanche, P.J., Waters, D., Shiboski, S., Rotter, J.I., Nickerson, D.A., Yang, H., Saad, M., *et al.* (2006). Phenotypic predictors of response to simvastatin therapy among African-Americans and Caucasians: the Cholesterol and Pharmacogenetics (CAP) Study. *Am J Cardiol 97*, 843-850.

Teslovich, T.M., Musunuru, K., Smith, A.V., Edmondson, A.C., Stylianou, I.M., Koseki, M., Pirruccello, J.P., Ripatti, S., Chasman, D.I., Willer, C.J., *et al.* (2010). Biological, clinical and population relevance of 95 loci for blood lipids. *Nature 466*, 707-713.

Thompson, P.D., Clarkson, P., and Karas, R.H. (2003). Statin-associated myopathy. *JAMA 289*, 1681-1690.

Tian, C., Gregersen, P.K., and Seldin, M.F. (2008). Accounting for ancestry: population substructure and genome-wide association studies. *Hum Mol Genet 17*, R143-150.

Tuokko, H., Frerichs, R., Graham, J., Rockwood, K., Kristjansson, B., Fisk, J., Bergman, H., Kozma, A., and McDowell, I. (2003). Five-year follow-up of cognitive impairment with no dementia. *Arch Neurol 60*, 577-582.

Voight, B.F., Kang, H.M., Ding, J., Palmer, C.D., Sidore, C., Chines, P.S., Burtt, N.P., Fuchsberger, C., Li, Y., Erdmann, J., *et al.* (2012). The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS Genet 8*, e1002793.

Voora, D., and Ginsburg, G.S. (2012). Clinical application of cardiovascular pharmacogenetics. *J Am Coll Cardiol 60*, 9-20.

Welch, C.L. (2012). Beyond genome-wide association studies: the usefulness of mouse genetics in understanding the complex etiology of atherosclerosis. *Arterioscler Thromb Vasc Biol 32*, 207-215.

Wilke, R.A., Ramsey, L.B., Johnson, S.G., Maxwell, W.D., McLeod, H.L., Voora, D., Krauss, R.M., Roden, D.M., Feng, Q., Cooper-Dehoff, R.M., *et al.* (2012). The clinical pharmacogenomics implementation consortium: CPIC guideline for SLCO1B1 and simvastatin-induced myopathy. *Clin Pharmacol Ther 92*, 112-117.

Willer, C.J., and Mohlke, K.L. (2012). Finding genes and variants for lipid levels after genome-wide association analysis. *Curr Opin Lipidol 23*, 98-103.

Zuk, O., Hechter, E., Sunyaev, S.R., and Lander, E.S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc Natl Acad Sci U S A 109*, 1193-1198.

APPENDICES

A-1. University of Western Ontario ethics approval

Research Western

Use of Human Participants - Ethics Approval Notice

Principal Investigator: Dr. Robert Hegele
Review Number: 07920E
Review Level: Delegated
Approved Local Adult Participants: 1840
Approved Local Minor Participants: 0
Protocol Title: Candidate gene sequencing, genetic and genomic analysis for identification of new genetic determinants of intermediate traits of atherosclerosis, dyslipidemia, diabetes, obesity, hypertension, lipodystrophy and other rare metabolic or cardiovascular disorders in the human population.
Department & Institution: Vascular Biology, Robarts Research Institute
Sponsor: Heart and Stroke Foundation of Canada
Canadian Institutes of Health Research

Ethics Approval Date: August 25, 2011 **Expiry Date:** December 31, 2015
Documents Reviewed & Approved & Documents Received for Information:

Document Name	Comments	Version Date
Revised UWO Protocol	Updated Sponsor Information	

This is to notify you that The University of Western Ontario Research Ethics Board for Health Sciences Research Involving Human Subjects (HSREB) which is organized and operates according to the Tri-Council Policy Statement: Ethical Conduct of Research Involving Humans and the Health Canada/ICH Good Clinical Practice Practices: Consolidated Guidelines; and the applicable laws and regulations of Ontario has reviewed and granted approval to the above referenced revision(s) or amendment(s) on the approval date noted above. The membership of this REB also complies with the membership requirements for REB's as defined in Division 5 of the Food and Drug Regulations.

The ethics approval for this study shall remain valid until the expiry date noted above assuming timely and acceptable responses to the HSREB's periodic requests for surveillance and monitoring information. If you require an updated approval notice prior to that time you must request it using the UWO Updated Approval Request Form.

Members of the HSREB who are named as investigators in research studies, or declare a conflict of interest, do not participate in discussion related to, nor vote on, such studies when they are presented to the HSREB.

The Chair of the HSREB is Dr. Joseph Gilbert. The UWO HSREB is registered with the U.S. Department of Health & Human Services under the IRB registration number IRB 00000940.

Ethics Officer to Contact for Further Information

X Janice Sutherland	Grace Kelly	Shantel Walcott

This is an official document. Please retain the original in your files.

A-2. Copyright permissions

1. **Dube, J.B.**, and Hegele, R.A. (2012). Genetics 100 for cardiologists: basics of genome-wide association studies. *Can J Cardiol 29*, 10-17.

ELSEVIER LICENSE
TERMS AND CONDITIONS

Mar 24, 2013

This is a License Agreement between Joseph B Dube ("You") and Elsevier ("Elsevier") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Elsevier, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

Supplier	Elsevier Limited The Boulevard,Langford Lane Kidlington,Oxford,OX5 1GB,UK
Registered Company Number	1982084
Customer name	Joseph B Dube
Customer address	
License number	3115600846797
License date	Mar 24, 2013
Licensed content publisher	Elsevier
Licensed content publication	Canadian Journal of Cardiology
Licensed content title	Genetics 100 for Cardiologists: Basics of Genome-Wide Association Studies
Licensed content author	Joseph B. Dubé,Robert A. Hegele
Licensed content date	January 2013
Licensed content volume number	29
Licensed content issue number	1
Number of pages	8
Start Page	10
End Page	17
Type of Use	reuse in a thesis/dissertation
Portion	figures/tables/illustrations
Number of figures/tables /illustrations	1
Format	print
Are you the author of this Elsevier article?	Yes
Will you be translating?	No
Order reference number	
Title of your thesis/dissertation	Genetic approaches to studying complex disease
Expected completion date	Jun 2013
Estimated size (number of pages)	120
Elsevier VAT number	GB 494 6272 12

2. **Dubé, J.B**., Johansen, C.T., Robinson, J.F., Lindsay, J., Hachinski, V., and Hegele, R.A. (2013). Genetic determinants of "cognitive impairment, no dementia". *J Alzheimers Dis 33*, 831-840.

Permission to publish manuscript in thesis

Carry Koolbergen Thu, Oct 11, 2012 at 9:02 AM
To: Joseph Brenton Dube

Dear Joseph Brenton Dube,

We hereby grant you permission to reproduce the below mentioned material in **print and electronic format** at no charge subject to the following conditions:

1. If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source, permission must also be sought from that source. If such permission is not obtained then that material may not be included in your publication/copies.

2. Suitable acknowledgement to the source must be made, either as a footnote or in a reference list at the end of your publication, as follows:

"Reprinted from Publication title, Vol number, Author(s), Title of article, Pages No., Copyright (Year), with permission from IOS Press".

3. This permission is granted for non-exclusive world **English** rights only. For other languages please reapply separately for each one required.

4. Reproduction of this material is confined to the purpose for which permission is hereby given.

Yours sincerely

Carry Koolbergen (Mrs.)

Contracts, Rights & Permissions Coordinator

Not in the office on Wednesday's

IOS Press BV

JOSEPH BRENTON DUBÉ

EDUCATION

2010-2013: **MSc Candidate (Completion June 2013)**

"Genetic characterization of Mendelian and complex vascular diseases"

Department of Biochemistry, Schulich School of Medicine & Dentistry, Robarts Research Institute, Western University, London, ON Canada

Canadian Institutes of Health Research (CIHR), Fellow in Vascular Biology

2006-2010: **BMSc Honours Specialization in Biochemistry**

"Post-translational modification of IGFBP1 in fetal growth restriction"

Department of Biochemistry, Schulich School of Medicine & Dentistry, Western University, London, ON Canada

AWARDS AND DISTINCTION

2011-2012: Queen Elizabeth II Graduate Scholarship ($15,000)

2011: Keystone Symposia travel scholarship ($1200)

2011: Schulich School of Medicine & Dentistry Research Award ($510)

2011: Best poster presentation, Taylor International Prize in Medicine

2010-2012: Western Graduate Research Scholarship (2 X $7,000)

2010-2012: CIHR and HSFC Vascular Research Training Program (2 X $6,000)

2010-2012: Heart and Stroke Foundation of Ontario Program Grant ($3,300)

2008-2010: Dean's Honor Roll

2008: Research Award, Schulich School of Medicine & Dentistry ($4,500)

RESEARCH POSITIONS

2010-2013.: **MSc Supervisor: Dr. Robert A. Hegele, Cardiovascular Genetics**

Robarts Research Institute, Western University, London, ON

- Designed and executed a two-stage genome-wide association study of pre-dementia using the Illumina high-density SNP genotyping CardioMetabochip in collaboration with the Broad Institute.
- Performed specialized genetic analyses including multi-dimensional scaling, multiple regression, risk score and mutation accumulation analyses using tools such as UNIX, PLINK, SAS, the UCSC genome browser and the HapMap dataset.
- Tested for replication of GWAS-identified signals within First Nations populations using TaqMan genotyping and Sanger sequencing.
- Tested for association between private *LDLR* mutations and lipid traits in First Nations using multiple linear regression and assessed the biological effects of these mutations on LDLR activity using *in vitro* assays in the CHO cell line.
- Supported lab activity by writing scientific reviews, reviewing research manuscripts within the peer review process, advising study design and statistical practice and supervising undergraduate student projects.

2008-2010: **BMSc Thesis Supervisor: Dr. M.B. Gupta, Fetal Growth Restriction**

Department of Paediatrics and Biochemistry, Schulich School of Medicine& Dentistry, Western University, London, ON Canada

- Investigated differential expression and phosphorylation of IGF binding protein-1 under hypoxic *in vitro* conditions using 1D- and 2D-PAGE.
- Designed experiments using Western blotting and native gel electrophoresis techniques to visualize protein expression and net electrical charge on protein isoforms.
- Conceptualized intracellular pathways regulating IGF binding protein-1 expression and phosphorylation as well as future experiments using siRNA-based transient knockdowns in *in vitro* models.

PREVIOUS WORK/LEADERSHIP/VOLUNTEER EXPERIENCE

2009-2010:	Summer house league coach, North York Cosmos Soccer Club
2008:	Summer intern, Ontario Ministry of Education
2007:	Summer intern, Ontario Ministry of Health and Long-Term Care
2005-2006:	Summer paralegal assistant, Idealogic Searchouse Corp.

PEER REVIEWED PUBLICATIONS

1. Fu J, Kwok S, Sinai L, Abdel-Razek O, Babula J, Chen D, Farago E, Fernandopulle N, Leith S, Loyzer M, Lu C, Malkani N, Morris N, Schmidt M, Stringer R, Whitehead H, Ban MR, **Dubé JB**, et al. (2013). Western Database of Lipid Variants (WDLV): A Catalogue of Genetic Variants in Monogenic Dyslipidemias. **In press**. (PMID: 23623477)
2. **Dubé, JB**, Johansen, CT, Robinson, J, Lindsay, J, Hachinski, V, Hegele, RA. Genetic determinants of "cognitive impairment, no dementia". *J Alzheimers Dis* 2013; 33(3):831-40. (PMID: 23042215)
3. **Dubé JB** and Hegele RA. Genetics 100 for cardiologists: basics of genome-wide association studies. *Can J Cardiol* 2013; 29(1):10-7. (PMID: 23200095)
4. **Dubé JB**, Hegele RA. The application of gene therapy in lipid disorders: where are we now? *Clin Lipidol*. 2012; 7(4):419-29**.**
5. **Dubé JB**, Boffa MB, Hegele RA, Koschinsky ML. Lipoprotein(a): more interesting than ever after 50 years. *Curr Opin Lipidol* 2012; 23(2):133-40. (PMID: 22327610)
6. **Dubé JB**, Johansen CT, Hegele RA. Sortilin: An unusual suspect in cholesterol metabolism. *Bioessays* 2011; 33(6):430-7. (PMID: 21462369)

PARTICIPATION IN PEER REVIEW PROCESS

Nature Genetics (2), Circulation (1), Circulation Research (2), Atherosclerosis (1)

ORAL AND POSTER ABSTRACT PUBLICATIONS

1. **Dubé, JB**, Johansen, CT, Robinson, J, Lindsay, J, Hachinski, V, and Hegele, RA. (2012). "Cardiovascular disease-related genetic variation in clinical cognitive impairment." **Oral** presentation at Western University's Department of Medicine Research Day in London, Ontario.
2. **Dubé, JB**, Johansen, CT, Robinson, J, Lindsay, J, Hachinski, V, and Hegele, RA. (2012). "Assessing the impact of cardiovascular disease-related genetic variation on cognitive impairment in neurodegeneration." **Poster** presentation at the 2012 Atherosclerosis, Thrombosis and Vascular Biology Scientific Sessions in Chicago, Illinois.
3. **Dubé, JB**, Johansen, CT, Robinson, J, Lindsay, J, Hachinski, V, and Hegele, RA. (2012). "Assessing the impact of cardiovascular disease-related genetic variation on

cognitive impairment in neurodegeneration." **Poster** presentation at London Health Research Day in London, Ontario.

4. **Dubé, JB**, Johansen, CT, Robinson, J, Lindsay, J, Hachinski, V, and Hegele, RA. (2012). "Assessing the impact of cardiovascular disease-related genetic variation on cognitive impairment in neurodegeneration." **Poster** presentation at the Keystone Symposium on ApoE, Alzheimer's and Lipoprotein Biology in Keystone, Colorado.
5. **Dubé, JB**, Johansen, CT, Robinson, J, Lindsay, J, Hachinski, V, and Hegele, RA. (2011). "Assessing the impact of cardiovascular disease-related genetic variation on cognitive impairment in neurodegeneration." **Oral** presentation given at the 36th Annual Canadian Lipoprotein Conference in Halifax, Nova Scotia.
6. **Dubé, JB**, Johansen, CT, Wang, J, Cao, H, Nykjaer, A, and Hegele, RA. (2011). "Rare genetic variation in *SORL1* is not associated with increased susceptibility to hypertriglyceridemia." **Poster** presented at the 2011 Atherosclerosis, Thrombosis and Vascular Biology Scientific Sessions in Chicago, Illinois.
7. **Dubé, JB**, Johansen, CT, Wang, J, Cao, H, Nykjaer, A, and Hegele, RA. (2011). "Rare genetic variation in *SORL1* is not associated with increased susceptibility to hypertriglyceridemia." **Poster** presented at Lawson Research Day in London, Ontario.
8. **Dubé, JB**, Johansen, CT, Wang, J, Cao, H, Nykjaer, A, and Hegele, RA. (2011). "Rare genetic variation in *SORL1* is not associated with increased susceptibility to hypertriglyceridemia." **Poster** presented at Western University's Margaret Moffat Graduate Research Day in London, Ontario.
9. **Dubé, JB**, Johansen, CT, Wang, J, Cao, H, Nykjaer, A, and Hegele, RA. (2011). "Rare genetic variation in *SORL1* is not associated with increased susceptibility to hypertriglyceridemia." **Poster** presented at Western University's Department of Medicine Research Day in London, Ontario.
10. **Dubé, JB**, Johansen, CT, Robinson, J, Lindsay, J, Hachinski, V, and Hegele, RA. (2010). "High throughput genome-wide assessment of cognitive impairment: study design." **Poster** presented at the 2010 J. Allyn Taylor International Prize in Medicine Symposium in London, Ontario.