RARE GENETIC VARIATION IN TYPE 2 DIABETES

LEVERAGING PUBLIC EXOME SEQUENCING DATA TO FIND RARE CAUSAL VARIANTS IN TYPE 2 DIABETES

BY JAMES FEINER, B.Sc., OCAD

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

Of the Requirements for the Degree Master of Science

MASTER OF SCIENCE (2021)	McMaster University
(Medical Sciences)	Hamilton, Ontario

TITLE:	Leveraging Public Exome Sequencing Data to Find Rare Causal Variants in Type 2 Diabetes
AUTHOR:	James Feiner, B.Sc., OCAD (University of Waterloo, 2012 & Seneca Collage, 2016)
SUPERVISOR:	Guillaume Paré, MD, M.Sc.
NUMBER OF PAGES:	102

ABSTRACT

Background: Type 2 Diabetes (T2D) is growing in prevalence worldwide over the last century. T2D incidence is linked to numerous complications, increased risk of heart disease, and oncology outcomes. This highlights the importance of preventive measures for T2D, wherein genetic predisposition can serve as an early warning sign. The role of rare variants (RVs) in T2D pathogenesis has not been adequately explored due to study size limitations, therefore we hypothesized that new associations could be found using publicly available data repositories.

Methods: Significant RV gene burden for T2D risk was discovered using exome sequences obtained from the United Kingdom Biobank (UKB) (n=162,215), then tested for replication in the Korean Association Resource project (n=973), the Metabolic Syndrome in Men Study (n=969), the San Antonio Mexican American Family Studies (n=309), and a pooled meta-analysis of the latter three cohorts. RV gene burden was reassessed in secondary analyses using T2D cases from each cohort and summary level data from the Genome Aggregation Database (GnomAD) (n=125,748).

Results: UKB exome wide significant associations were found in *GCK* (OR=2.44, p=8.91×10^{-11}) and *PAM* (OR=1.32, p=1.39×10^{-6}), and suggestive associations (p<0.001) were found in 33 additional genes. Replication was limited in KARE, METSIM, SAMAFS and in the secondary analyses with GnomAD because of limited sample sizes and miscalibration with the external control, respectively. Follow-up analyses include exploration of RV gene burden in additional diabetes subtypes, evaluation of clinical features between RV carriers and non-carriers, comparing the ability to predict T2D with rare variant, polygenic, and phenotypic risk scores. Methodological improvements include the incorporation of robust analytic tools and increasing access to a greater diversity and number of samples.

Conclusion: Publicly available exome sequencing data has identified genes where RV burden affects T2D pathogenesis and risk. The study of rare genetic variation in diabetes is just beginning.

ACKNOWLEDGEMENTS

I would like to foremost thank Dr. Guillaume Paré for his seemingly eternal patience and guidance throughout the long journey that has been my master's degree. I'd like to think that I've paid him back somewhat by appeasing his sweet tooth, but not even a lifetime supply of the *Feinest* baking would be enough thanks. I appreciate the feedback provided by committee members Dr. Zubin Punthakee and Dr. Hertzel, particularly when stressing the importance of communicating the "big picture" of my research. Also, thank you Dr. Marie Pigeyre for many explanations to help my understanding of diabetes, providing gene-sets, and for guiding me to several interesting directions of research.

Mike Chong and Ricky Lali were crucial in showing me the ropes of genomic bioinformatics and the science behind the methodologies. Both are responsible for building multiple resources and cleaning datasets used by many a student, myself included of which I can't thank them enough creating such wonderful steppingstones. I would rue the day that I'd be forced to choose allegiance between the two men, so hopefully that never comes to pass.

I eventually moved away from exome sequencing data generated locally at the genetic and molecular epidemiology laboratory, but I remember my roots. So, thank you Amanda Hodge, Shana Lamers, and Reina Ditta for training me when I was in the wet lab and then continue to run it long after I had dried off. Also, I could not have attempted my first project without the work of Taylor MacIsaac. Lastly, I appreciated that my food drops seemed

to please the ever-circulating tide of co-op students, undergraduates, and whomever else partook.

I appreciate the time spent with Loubna Ahkabir, Irfan Khan, Ann Le, Pedrum Mohammadi-Shemirani, Matteo Di Scipio, Tafadswa Machipisa, Nazia Pathan, Rob Morton, and Michael Chong slogging it out for many hours on the thesis writing group or answering, technical questions, and further thanks for the latter six individuals for reading and editing my thesis. Also, thanks to Harry Wang and Sukrit Narula for clarifying whether my parts of my background and discussion made clinical sense.

Spending time at the lab was also fun and rewarding, and to that end I'd like to further thank Mike Chong and Ricky Lali for their courtside mentorship, Michel Kiflen for his frequently entertaining hubris, and Nathan Cawte being a Legend.

Finally, I want to thank my family, in particular my parents Susan and Steve Feiner for their continued support and extensive advice throughout the last 4 years. My siblings and their kids were alright, too. While not the right species to directly benefit from the work, I would also like to dedicate this thesis to Howard. He was a good boy and passed away before his time.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AFR	African/African American
AC	Allele Count
AOD	Age of Diagnosis
AMR	American admixed/Latino
ASJ	Ashkenazi Jewish
ABCC8	ATP-binding Cassette, Subfamily C, Member 8
BAM	Binary Alignment Map
BMI	Body Mass Index
BED	Browser Extensible Data
CEL	Carboxyl Ester Lipase
CDCV	Common Disease, Common Variant Hypothesis
CDRV	Common Disease, Rare Variant Hypothesis
CAD	Coronary Artery Disease
CMAC	Cumulative Minor Allele Count
CMAF	Cumulative Minor Allele Frequency
dbGaP	Database of Genotypes and Phenotypes
dNTP	Deoxynucleotide Triphosphate
DNA	Deoxyribonucleic Acid
DM	Diabetes Mellitus
EAS	East Asian
EGA	European Genome-phenome Archive
FPG	Fasting Plasma Glucose
FIN	Finnish
GOF	Gain of Function
gCF	Gene Correction Factor
GnomAD	Genome Aggregation Database

GRCh37	Genome Reference Consortium Human Genome Build 37
GRCh38	Genome Reference Consortium Human Genome Build 38
GWAS	Genome Wide Association Study
λ_{med}	Genomic Inflation Factor
GCK	Glucokinase
HWE	Hardy Weinberg Equilibrium
HbA1c	Hemoglobin A1C
HNF1A	Hepatocyte Nuclear Factor-1-alpha
HNF4A	Hepatocyte Nuclear Factor-4-alpha
HGMD	Human Gene Mutation Database
HGP	Human Genome Project,
iCF	Individual Correction Factor
ICD-10	International Classification of Diseases, Tenth Revision
KMS	KARE-METSIM-SAMAFS Meta-analysis
KARE	Korea Association Research
LDL	Low Density Lipoprotein
LD	Linkage Disequilibrium
LOF	Loss of Function
MODY	Mature Onset Diabetes of the Young
MCAP	Mendelian Clinically Applicable Pathogenicity Score
METSIM	Metabolic Syndrome in Men Study
MAC	Minor Allele Count
MAF	Minor Allele Frequency
NFE	non-Finnish European
OR	Odds Ratio
PAM	Peptidylgycine α-amidating Monooxygenase
PCR	Polymerase Chain Reaction
PC	Principal Component

BLK	Proto-Oncogene, Src Family Tyrosine Kinase
P	P-value
QQ	Quantile Quantile
RV	Rare Variant
RVGRS	Rare Variant Genetic Risk Score
RV-EXCALIBER	Rare Variant Exome CALIBration using External Repositories
DP	Read Depth
SPA	Saddle Point Approximation
SAMAFS	San Antonio Mexican American Family Studies
SKAT	Sequence Kernel Association Test
SNP	Single Nucleotide Polymorphism
dbSNP	Single Nucleotide Polymorphism Database
SNV	Single Nucleotide Variant
SAS	South Asian
TOPMed	Trans-Omics for Precision Medicine
T1D	Type 1 Diabetes
T2D	Type 2 Diabetes
UKB	UK Biobank
VCF	Variant Call Format
WGS	Whole Genome Sequencing

CHAPTER 1 – Background

1.1 Overview

This chapter provides background information on two topics integral to this thesis: diabetes and the genetics of disease.

1.2 Diabetes Mellitus

1.2.1 History and Physiology

Despite its presence throughout human history, the nature of the chronic disease Diabetes Mellitus (DM) had remained a mystery for centuries. Records from ancient Egypt and India described DM by intense thirst and excessive production of urine with a characteristically sweet taste [1]. Chinese, Greek, and Arab physicians in the early Common Era observed additional long-term symptoms like deteriorated eyesight, frequent skin infections, and emaciation [1]. Only starting in the 18th and 19th centuries did various scientists sequentially uncover the physiological nature of DM. Breakthroughs, made primarily through animal experimentation on canines, included: the discovery of a substance ("glycogen") in the liver that raised blood sugar (glucose); that removal of the pancreas resulted in sugary urine ("glycosuria"); and that the diabetic state was reversed after intravenous injection of a purified pancreas extract ("insulin") [1]. This accumulated knowledge has facilitated the modern understanding of DM.

Blood glucose concentration is tightly regulated in humans to be within 4 to 6 mmol/L, achieved through the endocrine hormone system known as glucose homeostasis (Figure 1.1) [2]. Glucose homeostasis can be briefly described as the interplay of insulin and glucagon. Insulin is created by β-cells in the pancreas and secreted in response to high blood glucose [2]. Endogenous insulin lowers blood glucose by stimulating the uptake of glucose from the blood into cells in adipose and muscle tissue, reducing the production of glucose from non-carbohydrates (gluconeogenesis), and increasing the conversion of glucose to glycogen that is stored in the liver (glycogenesis) [2]. Conversely, glucagon is created by alpha-cells in the pancreas in response to low blood glucose, promoting gluconeogenesis and conversion of glycogen back into glucose (glycogenolysis) [2]. DM is the manifestation of this system failing: where insufficient production of insulin, often coupled with its reduced efficacy, causes chronic high blood glucose (hyperglycemia).

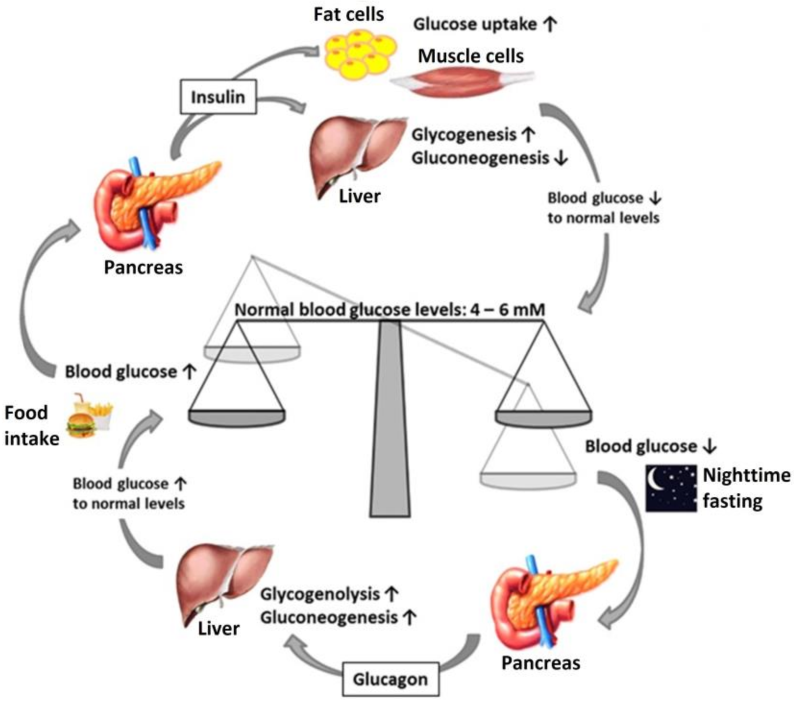


Figure 1.1: Generalized overview of glucose homeostasis. While the roles of insulin and glucagon are highlighted here, the system involves several additional hormones that effect a multitude of factors including satiety, digestion, β-cell proliferation, and secretion regulation (Adapted from Röder, et al., 2016) [2].

1.2.2 Prevalence, Complications, and Classification

Within the span of the last couple hundred years, DM had gone from a rare affliction of antiquity to a global epidemic [3]. Throughout most of human civilization, life expectancy was lowered due to widespread outbreaks of infectious disease and general malnutrition [3,4]. In western countries this started to change halfway through the 19th century when public health initiatives, improved standards of living, and industrialized agriculture were adapted [5]. However, as these populations lived longer, chronic diseases like DM and obesity began to manifest at an ever-increasing rate [3]. The prevailing early 20th century explanation for this phenomenon was that decreased physical activity and increased consumption of sugary food led to obesity, of which DM was a complication [3]. While diet and lack of exercise continue to be recognized as DM risk factors, the contemporary view is that socioeconomic status strongly dictates adherence to a healthy lifestyle and access to medical care [6]. Without this nuanced understanding, early public health initiatives were unsuccessful in preventing the epidemic from growing over time [3]. For example, the number of adults with diabetes in 2000 was estimated globally at 151 million [9]. By 2019 this increased over three fold to an estimated 463 million, equivalent to a worldwide prevalence of over 9% [9]. While DM can be managed through medical or lifestyle interventions, inequity in medical access and education leaves many cases undiagnosed or untreated [10]. This is evidenced by developing countries currently contributing to 80% of DM cases [9,10].

DM is dangerous because hyperglycemia damages capillaries leading to microvascular and macrovascular complications. For example, vison loss in diabetic retinopathy occurs because damage to the microvasculature of the retina causes hypoxia from reduced blood flow, leakage from ruptured blood vessels, and abnormal capillary proliferation [11,60]. The association of hyperglycemia with atherosclerosis in the macrovasculature may be due to capillary damage in the vasa vasorum [60]. Alternatively, it may be due to the build-up of cholesterol plaque that narrows the arteries, reducing blood flow to the heart or to other peripheral organs and tissues [11]. Furthermore, plaques can rupture and form blood clots that can cause myocardial infarctions or strokes by blocking blood flow to the heart or brain, respectfully [11]. People with DM are up to 4 times more likely to develop cardiovascular disease, which lead as causes of death for the chronic disease [12]. DM can also induce cell proliferation in carcinogenesis, possibly leading to cancers [13]. The risk of all-cause mortality for people with DM is 50% greater than those without, and each year DM accounts for millions of deaths worldwide [8].

Several different etiologies can lead to DM, so the disease has been classified into specific subtypes. The two most commonly occurring are Type 1 Diabetes (T1D) and Type 2 Diabetes (T2D). T1D occurs in less than 10% of DM cases, and mostly in juveniles, where production of insulin is hindered due an autoimmune reaction that destroys β-cells [14]. While risk for T1D has traditionally been explained by genetic predisposition, there is now evidence that environmental factors can help trigger the autoimmune condition [15]. By a large margin, the predominant form of DM is T2D, with a case prevalence of over 90% in adults and between 20% to 45% in certain populations of children [16]. It is notable that the age of onset for T2D has been decreasing over time, as prior to the mid 1990's the case prevalence in children was only 1 to 2% [16]. T2D has traditionally been characterized by resistance to insulin and a subsequent inadequate production of the hormone, though efforts to reclassification efforts [14,63]. Several hypotheses explaining the mechanisms behind insulin resistance have been proposed, though their clinical application has been limited [17]. In contrast, risk factors for T2D are well documented and are frequently referenced in both preventive and diagnostic settings. Environmental risk factors include an overweight body mass index (BMI), low activity levels, and poor socioeconomic status [16]. Inherited risk factors include prior family history of T2D, as well as belonging to an ethnic minority which have higher incidence rates [16]. These risk factors are invaluable early warning signs to delay or slow the progression of T2D.

1.2.3 Diagnosis, Treatment, and Prevention

Diagnosis of T2D, as well as evaluating efficacy of treatments, can be achieved with rapid tests of blood glucose concentration such as the fasting plasma glucose test (FPG). Long-term glycemic control over a 3-month period can be monitored by measuring levels of glycated hemoglobin (HbA1C) [61]. Results from either test can be used to diagnose a normal, prediabetic or diabetic state (Table 1.1) [18]. While not guaranteeing progression to diabetes even in the absence of interventions, prediabetes nevertheless incurs a greater risk of diabetes and associated complications [19,61]. Furthermore, if the diabetic state progresses past the diagnostic threshold, the chance of morbidity and mortality increases [20].

Table 1.1: Diagnostic ranges for T2D & interventions

Diagnosis	**HbA1C (%)**	**FPG (mmol/L)**	**Interventions**
Normal	< 6	< 6.1	Healthy lifestyle promotion, risk factor assessment
Prediabetes	6 – 6.4	6.1 – 6.0	Lifestyle changes, Metformin, risk factor assessment
Diabetes	≥ 6.5	≥ 7.0	Lifestyle changes, Metformin, Second-line medications

Two standard clinical tests and thresholds used to diagnose T2D, as well as interventions used to prevent disease progression.

Progression of the diabetic state is gradual and can be interrupted, and possibly reversed, with lifestyle changes and medical interventions. Exercise increases glucose uptake and glucose metabolism in skeletal muscles [21,22]. Weight loss achieved through dieting reduces adiposity and fat distribution, improving insulin regulation [23]. The first-line medication for T2D treatment is metformin, which increases insulin sensitivity [24,25]. There are also several second-line drugs that also improve insulin sensitivity or increase insulin secretion, as well as insulin therapy [25].

Although T2D has many treatment options, the earlier the detection of the chronic disease, the better the prognosis [26]. In asymptomatic or prediabetic people, environmental and inherited risk factors can be assessed to estimate the probability of developing T2D [27]. For example, overweight individuals have at least 50% greater risk of getting T2D than those with normal BMI [28]. Prior family history has been estimated to convey a 40% or 70% lifetime risk, depending on whether one or both parents had T2D, respectfully [29]. While serving as valuable pre-diagnostic and monitoring tools, these conventional risk factors are hindered by confounding effects and imprecise associations [30]. However, improvements in genetics over the last two decades have made it usable for early detection of disease, as well as potentially provide pathophysiological insights, identify novel drug targets, or find sub-populations of patients.

1.3 *Disease Genetics*

1.3.1 Mendelian versus Complex Disease

One of the challenges with early genetics was figuring out how chronic diseases like T2D fitted within Mendelian patterns of inheritance. In the 19th century, the Austrian scientist Gregor Mendel discovered that traits are inherited though genes, with the paternal and maternal sides each contributing one version (allele) of each gene [31]. Specific allele pairings (genotypes) correspond to variations of a trait (phenotypes), where the expressed phenotype corresponds to the dominant allele in the genotype [31]. Mendel also postulated that each gene is inherited independently of other genes [31]. One example of a Mendelian, or monogenic, disease is Mature Onset Diabetes of the Young (MODY), a rare autosomal dominant form of DM present in up to 5% of cases (Figure 1.2, Table 1.2) [32]. However, most chronic diseases seemed to defy Mendelian patterns of inheritance because they were observed to have continuous phenotypic expression, such as varying age of onset in T2D [33]. Over time it became clear that these diseases were multifactorial with both environmental and polygenic components, with the latter referring to the cumulative effect of multiple mutations across multiple genes. A new model was required to explain such complex patterns of inheritance [33].

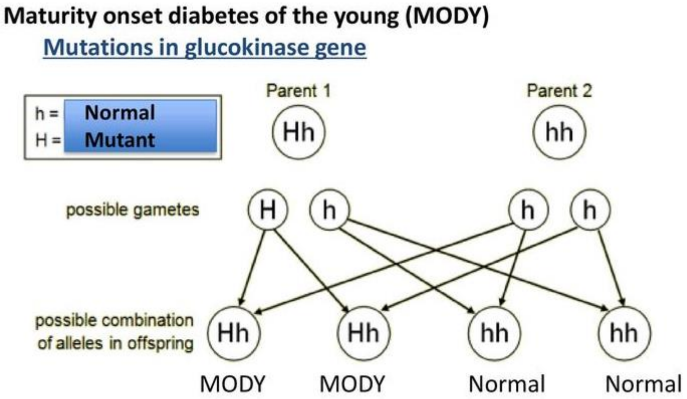

Maturity onset diabetes of the young (MODY)
Mutations in glucokinase gene
h = Normal
H = Mutant
Parent 1
Hh
Parent 2
hh
possible gametes
H
h
h
h
possible combination of alleles in offspring
Hh
Hh
hh
hh
MODY
MODY
Normal
Normal

Figure 1.2: Mendelian inheritance of MODY. Diploid parental germ cells undergo meiosis and form haploid gametes, each with one allele per gene. Fertilization between male and female gametes results in diploid offspring, where two alleles combine as a genotype. The MODY phenotype is expressed if the glucokinase genotype contains at least one dominant mutant allele, which was the case for Parent 1 and had a 50% chance of occurring in the offspring [34].

Table 1.2: Overview of MODY genes

Gene	**Percent of MODY cases**	**Pathophysiology**	**Microvascular complications**
HNF1A	30–65	β-cell dysfunction: mainly insulin secretory defect	Common
GCK	30–50	β-cell dysfunction: glucose-sensing defect	Rare
HNF4A	5–10	β-cell dysfunction: mainly insulin secretory defect	Common
HNF1B	<5	β-cell dysfunction	Common
PDX1	1	β-cell dysfunction	Unknown
ABCC8	<1	ATP-sensitive potassium channel dysfunction	Unknown
APPL1	<1	Insulin secretion defect	Unknown
BLK	<1	Insulin secretion defect	Unknown
CEL	<1	Pancreatic endocrine & exocrine dysfunction	Unknown
INS	<1	β-cell dysfunction	Unknown
KCNJ11	<1	ATP-sensitive potassium channel dysfunction	Unknown
KLF11	<1	Decreased glucose sensitivity of β-cells	Unknown
NEUROD1	<1	β-cell dysfunction	Unknown
PAX4	<1	β-cell dysfunction	Unknown

Adapted from Naylor, Knight, and del Gaudio, 2018 [57].

The prevailing hypothesis explaining inheritance of complex disease is based on the interaction of allele frequency and penetrance. Alleles that cause or contribute to disease are subject to negative selective pressure, which over successive generations should result in their lowered population frequency [33]. Certain alleles may protect against disease pathogenesis and have an increased population frequency due to positive selective pressure [33]. However, an allele may have pathogenic, protective, or benign effects on different disease phenotypes, each influencing its population frequency [33]. At any given genomic location, there is one major allele that is most prevalent and any number of minor alleles with lower prevalence. Therefore, the term minor allele frequency (MAF) is used to denote allele population prevalence.

The effect size or penetrance of an allele will vary depending on its functional impact, which can range from negligible to intermediate to singlehandedly causing or preventing disease [35]. As is the case with MODY and most Mendelian disease, alleles of high penetrance that cause the disease phenotype are strongly selected against and have low to rare MAFs [36]. In contrast, the higher prevalence of complex diseases indicates that the alleles contributing to their pathogenesis are better tolerated, allowing them in turn to retain higher MAFs [36]. This is the basis of the common disease, common variant (CDCV) hypothesis (Figure 1.3), which has pervaded the field of modern genomics [33].

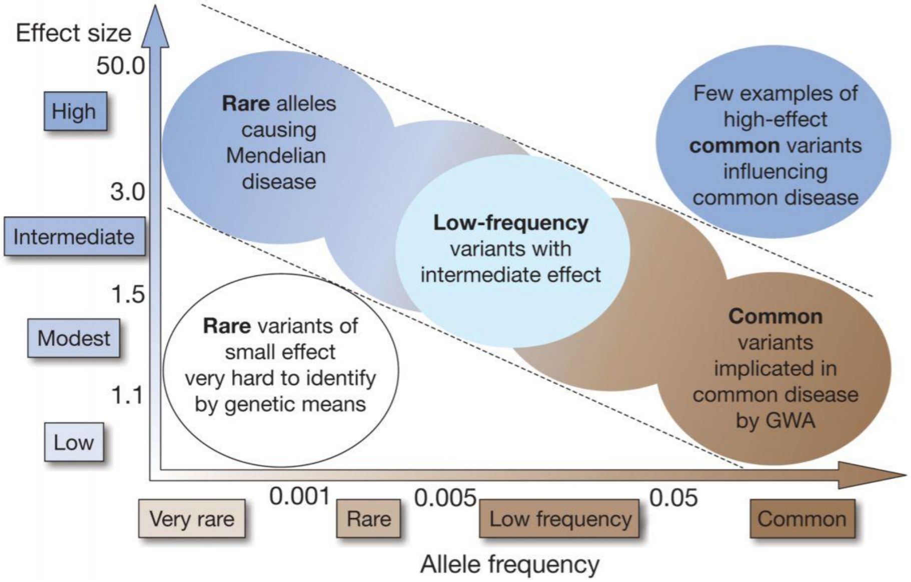

Effect size
50.0
High
3.0
Intermediate
1.5
Modest
1.1
Low
Rare alleles causing Mendelian disease
Low-frequency variants with intermediate effect
Few examples of high-effect common variants influencing common disease
Rare variants of small effect very hard to identify by genetic means
Common variants implicated in common disease by GWA
Very rare
0.001
Rare
0.005
Low frequency
0.05
Common
Allele frequency

Figure 1.3: Common disease, common variant hypothesis. Effect size on the y-axis represented as odds ratios, which compares disease prevalence among carriers of the alleles and non-carriers [36]. Variants are different versions of alleles at specified genomic locations (position on a chromosome).

1.3.2 GWAS and Missing Heritability

By the 1990's research in T2D genetics was making steady progress with the identification of associated genes through linkage analysis and targeted sequencing [29]. Though at the time there were limitations for these techniques: linkage analysis offered low resolution of the genome and performed poorly in identifying polygenic alleles, whereas the cost and time requirements of targeted sequencing restricted studies to sample sizes unrepresentative of general populations [29,37]. Both issues would be addressed with the completion of the Human Genome Project (HGP) in 2003, a colossal project that took 13 years of international collaboration and cost – adjusted for inflation – upwards of $5 billion [37]. In addition to creating the first full sequence of the human genome, the HGP also expedited development of technology and software that would transform the field of genomics [38]. While advancements coincided with a massive reduction in the price of genomic laboratory techniques, in 2006 whole genome sequencing remained prohibitive by costing tens of millions of dollars per person [37]. However, by then single nucleotide polymorphism (SNP) based arrays were becoming the predominant and cost-effective means of conducting a new kind of genetic analysis, the genome wide association study (GWAS) [39].

Over the last 15 years, GWAS have been the major driver in the discovery of variants associated with T2D and other complex disease [40]. In a GWAS, the frequency of variants spanning the genome are compared between cases and controls of a disease phenotype [39]. Variants associated with the disease are those with significantly different MAFs between the groups and can be pathogenic or protective if there is a higher or lower frequency in cases, respectfully [39]. GWAS usually employ genotyping arrays, which contain probes corresponding to SNPs across the genome. Deoxyribonucleic acid (DNA) is extracted, exposed to the genotyping array, and any probe that hybridizes to the DNA generates a signal indicating presence of the given SNP in the sample [39]. This process is easily automated and has allowed high throughput GWAS with tens and hundreds of thousands of samples [40]. Multiple GWAS have investigated T2D and to date more than 500 associated variants have been identified (Figure 1.4) [41]. These findings have provided insights into T2D pathology and have helped established a strong basis for polygenic risk scores, where the cumulative effect of the variants is used as a measure of genetic predisposition in an individual [42]. However, these variants have an average small effect size and collectively they only explain 19% of T2D heritability, a fraction of the 72% estimated from twin studies [41,62]. This phenomenon, known as “missing heritability”, has been characterized in other complex disease and suggests that the CDCV hypothesis is insufficient on its own.

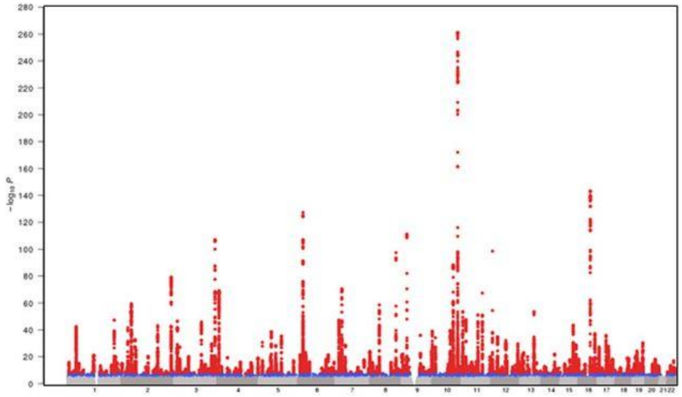


Figure 1.4: Manhattan plot of T2D GWAS associations. Adapted from the meta-analysis by Vujkovic, et al., 2020. Shown here are results from the European subset of 148,726 T2D cases and 965,732 controls. Red points are variants that achieved genome wide significance ($p < 5.0\times10^{-8}$), their chromosomal locations approximated on the x-axis.

One explanation for missing heritability is the common disease, rare variant (CDRV) hypothesis. CDRV postulates that predisposition for complex disease is driven by high penetrance rare variants (RVs) that occurred too recently, approximately over the last two centuries, to be eliminated by natural selection [33]. However, the study of RVs on a population level has been a challenge with GWAS. Early SNP arrays only had a few thousand probes, necessitating the selection of probes that corresponded to commonly occurring variants to maximize representation of the genome [39,43]. Probe density on arrays has increased and the ability to infer uncalled genotypes through imputation has improved over time, yet SNP arrays still have limited capacity in detection of RVs with MAFs under 1% or 0.1% [41].The proper investigation of the role of RVs in complex disease required a return to sequencing.

1.3.3 Affordable Exome Sequencing

The HGP was largely completed using automated Sanger Sequencing, a very accurate methodology that is limited by high expenses, slow determination of base pairs, and low sample throughput [58]. Over the course of the early 2000s, next generation sequencing (NGS) technology was developed and solved or improved all these problems [58]. In NGS there is almost unlimited sensitivity for detection of RVs, though this technology still costs a lot more compared to SNP genotyping arrays. However, the nature of RVs means that expensive sequencing of the entire genome is not necessary. RVs of high penetrance are likely to be located within one of the approximately 20,000 protein coding genes [43]. The coding regions within genes are called exons, which collectively are called the exome and consist of 1–2% of the base pairs in the genome [43]. Consequently, exome sequencing was used as a less expensive and comprehensive alternative to whole genome sequencing to study RV associations [43]. In exome sequencing, exonic regions of DNA are selectively captured and replicated with the polymerase chain reaction (PCR). During PCR, addition of nucleotides generates a signal that allows simultaneous physical construction of the replicates and digital versions called “reads” [44]. The reads are then aligned and compared to a reference genome, where differences in the DNA sequence are identified as variants [44]. While there is no limit to the rarity of variants that can be detected with exome sequencing, the number of variants and ability to analyze them is dependant on the number of samples in a study [45].

Like in a GWAS, analysis of RV disease association involves comparison of variant MAFs between exome sequences of cases and controls. However, the low population frequency of RVs makes it unlikely that any one variant will be present in two or more samples in a study [43]. This inconsistency severely reduces the statical power of analysing individual RVs [43]. One solution is to instead compare the cumulative burden of all RVs in each gene, assuming there is a shared functional impact of RVs in each region [46]. Previous exome sequencing studies have used this RV gene burden technique, yet they have not been able to find many statistically significant RV associations with T2D [47]. One explanation is that exome sequencing studies require tens of thousands of samples to capture rare variation that is representative of the population, [45]. Currently, exome sequencing costs around $300 per person, about ten times the price of genotyping with a SNP array covering the full genome [37]. The financial cost of exome sequencing projects can become prohibitive when studies require several thousand samples [37]. To obtain the required sample sizes to analyze RVs, another option is use publicly available data.

1.3.4 Publicly Available Data

Since the international collaboration of the HGP, sharing of genomic data has continued worldwide [48]. Currently there are innumerable publicly available resources, ranging catalogues of GWAS results, genome reference builds, annotation databases, and open-source software. From these, a trove of information can be used for a well powered analysis of RV gene burden in T2D.

When approaching genomics on a global scale, it is important consider the ethnicity and ancestry of the exome sequenced samples. Historically, the distribution of genetic studies around the globe has not been uniform, with the majority conducted in seven distinct ancestry groups. These are categorized as five major and two minor super populations: non-

Finish European (NFE), East Asian (EAS), South Asian (SAS), American admixed/Latino (AMR), African/African American (AFR), Finnish (FIN), and Ashkenazi Jewish (ASJ) [50]. FIN and ASJ are separate from NFE because the two groups are genetically bottlenecked, and FIN is also overrepresented compared to the rest of the continent [49]. On its own, NFE contributes to over half of the exomes in publicly available exome sequencing data [50]. Relative to the remaining four super populations, AFR is the most under-represented [51]. This disparity is problematic because MAF of risk alleles can vary considerably between the groups and affect disease susceptibility [52]. Extrapolating genetic insights from an NFE cohort to another ethnic or ancestry group can be inaccurate and even dangerous when influencing clinical decisions [50]. There are a growing number of multi-ethnic exome sequencing studies being made publicly available, however inequality persists.

Table 1.3: Breakdown of T2D cases by ethnicity and ancestry in the UKB

Ethnicity	**T2D cases**	**Percent of total**	**All samples**	**Percent of total**
African	918	2.79	7504	1.56
British	27920	84.78	431102	89.42
Indian	988	3.00	7465	1.55
Non-British Caucasian	1679	5.10	28581	5.93
South Asian	1427	4.33	7465	1.55
Total	32932		482117	

One of the biggest and newest genomic resources consists of participants who are almost 90% of British ancestry (Table 1.3), though it is still an invaluable resource and cannot be discounted for its relative lack of diversity. The United Kingdom Biobank (UKB) is a massive conglomeration of patient information including diagnostics, clinical endpoints, laboratory measurements, and genomic data for over 500,000 participants [53].

Patient level data from the UKB is accessible to researchers upon application approval and monetary fees. SNP array-based genotyping and exome sequencing of all UKB participants have been completed, though so far only about 200,000 exomes have been released. With over 32,000 T2D cases in the full database, this resource can drive the discovery of RV associations to the complex disease.

There are some large, UKB-sized exome sequence initiatives for underrepresented populations that are underway, like the Human, Heredity, and Health in Africa project. In the meantime, available online are data from a multitude of smaller studies from diverse populations. One of the largest repositories is the Database of Genotypes and Phenotypes (DBGap) which contains over 240,000 exome sequences from its collected studies. Access to this data on DBGap is granted to researchers by an approval process and can be downloaded on a study-by-study basis. A previous meta-analysis used exome sequences from over 40,000 T2D cases and controls, primarily from DBGap (Table 1.4) [47]. This data could be accessed again and used to validate RV associations found in the UKB. Additionally, the small individual studies could be analyzed with another database that approximates the general population.

Table 2.4: Breakdown of 26 T2D studies available from DBGap

Study name	Main ancestry	T2D cases	T2D Controls
Wake Forest School of Medicine Study	AFR	518	532
Jackson Heart Study	AFR	502	527
BioMe Biobank Program	AFR	1294	1254
Exome Sequencing Project A	AFR	467	1374
Exome Sequencing Project B	NFE	390	2843
Korea Association Research Project	EAS	526	561
Korea Seoul National University Hospital	EAS	450	475
Research Studies in Hong Kong	EAS	493	485
Malmo-Botnia Study	FIN	478	443

UKT2D Consortium	NFE	322	320
Metabolic Syndrome in Men Study	FIN	484	498
Ashkenazi	ASJ	506	355
Genetics of Diabetes and Audit Research Tayside Study	NFE	960	966
Framingham Heart Study	NFE	396	596
Mexico City Diabetes Study	AMR	281	549
Multiethnic Cohort	AMR	1476	1443
Diabetes in Mexico Study	AMR	1522	1546
UNAM/INCMNSZ Diabetes Study	AMR	1998	1977
Starr County, Texas	AMR	1762	1738
San Antonio Mexican American Family Studies	AMR	272	218
Singapore Indian Eye Study	SAS	1640	1478
London Life Sciences Population Study	SAS	531	538
Pakistan Genomic Resource	SAS	914	932
SEARCH for Diabetes in Youth	Multi-ethnic	533	0
Finland-United States Investigation of NIDDM Genetics Study	FIN	472	476
Treatment Options for Type 2 Diabetes in Adolescents and Youth	Multi-ethnic	3097	0
Total		22284	22124

An alternative way of analysing RVs, especially those from smaller studies, is to utilize summary level data as external controls [54]. The Genome Aggregation Database (GnomAD) is a consortium providing summary level genotypes, with the current release consisting of 125,748 human exomes (Figure 1.5) [55]. Curation of GnomAD involved the removal of related individuals and cases of pediatric disease. These steps helped ensure that disease prevalence in GnomAD would not exceed that of the general population [56]. The large size of this database allows for the detection of rare alleles that may otherwise go undetected in a smaller study.

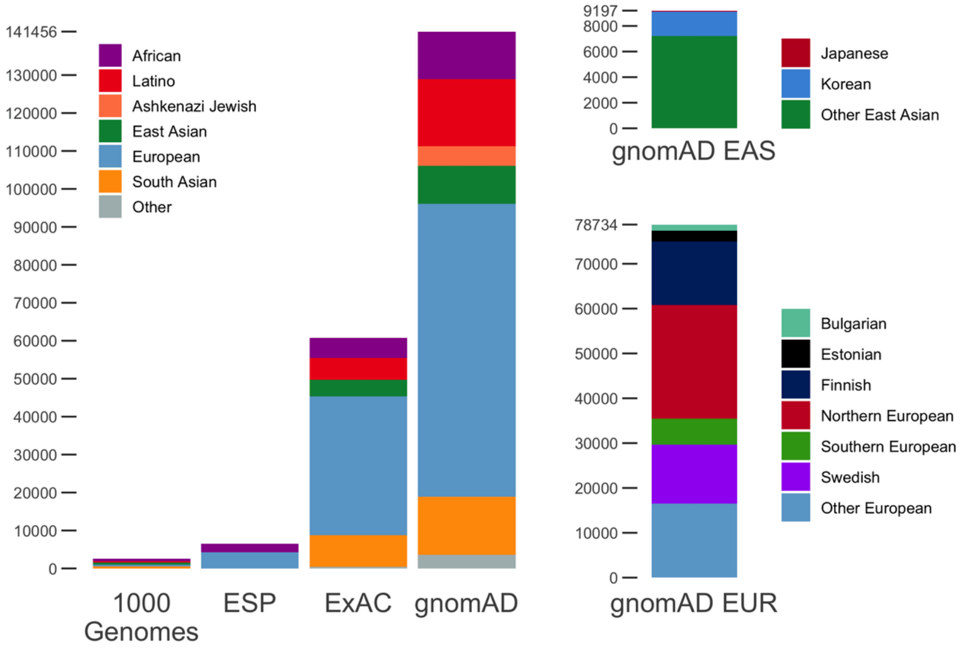


Figure 1.5: Breakdown of GnomAD superpopulations. Bar plots showcasing the ancestries and ethnicities of participants that contributed to GnomAD's exome sequences [55].

1.4 Summary

The chronic hyperglycemia indicative of DM has a well characterized pathology and there are several efficacious prevention and treatment strategies for the disease. However, global prevalence of DM has been rapidly growing, putting people with T2D at risk of severe complications like diabetic retinopathy, stroke, or heart disease. This accentuates the need for early detection tools that predict genetic predisposition for T2D. While the effects of common variation have been extensively explored, recent advancements in exome sequencing affordability and data availability invite RV analyses.

1.5 References

1. Karamanou, M., Protogerou, A., Tsoucalas, G., Androutsos, G., & Poulakou-Rebelakou, E. (2016). Milestones in the history of diabetes mellitus: The main contributors. *World journal of diabetes*, *7*(1), 1–7. https://doi.org/10.4239/wjd.v7.i1.1

2. Röder, P. V., Wu, B., Liu, Y., & Han, W. (2016). Pancreatic regulation of glucose homeostasis. *Experimental & molecular medicine*, *48*(3), e219. https://doi.org/10.1038/emm.2016.6

3. Karamanou, M. (2016). Milestones in the history of diabetes mellitus: The main contributors. *World Journal Of Diabetes*, *7*(1), 1. doi: 10.4239/wjd.v7.i1.1

4. Eknoyan, G. (2006). A History of Obesity, or How What Was Good Became Ugly and Then Bad. *Advances In Chronic Kidney Disease*, *13*(4), 421-427. doi: 10.1053/j.ackd.2006.07.002

5. Shaw-Taylor L. (2020). An introduction to the history of infectious diseases, epidemics and the early phases of the long-run decline in mortality. *The Economic history review*, *73*(3), E1–E19. https://doi.org/10.1111/ehr.13019

6. Baker, E. A., Schootman, M., Barnidge, E., & Kelly, C. (2006). The role of race and poverty in access to foods that enable individuals to adhere to dietary guidelines. *Preventing chronic disease*, *3*(3), A76.

7. Noymer, A., & Garenne, M. (2000). The 1918 influenza epidemic's effects on sex differentials in mortality in the United States. *Population and development review*, *26*(3), 565–581. https://doi.org/10.1111/j.1728-4457.2000.00565.x

8. Rowley, W. R., Bezold, C., Arikan, Y., Byrne, E., & Krohe, S. (2017). Diabetes 2030: Insights from Yesterday, Today, and Future Trends. *Population health management*, *20*(1), 6–12. https://doi.org/10.1089/pop.2015.0181

9. International Diabetes Federation. (2019). *IDF Diabetes Atlas, 9th edn.* Retrieved 24 September 2020, from: https://www.diabetesatlas.org

10. Tabish S. A. (2007). Is Diabetes Becoming the Biggest Epidemic of the Twenty-first Century?. *International journal of health sciences*, *1*(2), V–VIII.

11. Fowler, M. (2008). Microvascular and Macrovascular Complications of Diabetes. *Clinical Diabetes*, *26*(2), 77-82. doi: 10.2337/diaclin.26.2.77

12. Hanefeld, M., Monnier, L., Schnell, O. et al. (2016). Early Treatment with Basal Insulin Glargine in People with Type 2 Diabetes: Lessons from ORIGIN and Other Cardiovascular Trials. *Diabetes Therapy*, 7(2): 187-201. doi: https://doi.org/10.1007/s13300-016-0153-3

13. Xu, C. X., Zhu, H. H., & Zhu, Y. M. (2014). Diabetes and cancer: Associations, mechanisms, and implications for medical practice. *World journal of diabetes*, *5*(3), 372–380. https://doi.org/10.4239/wjd.v5.i3.372

14. Brutsaert, E. (2020). Diabetes Mellitus (DM) - Endocrine and Metabolic Disorders - MSD Manual Professional Edition. Retrieved 8 March 2021, from https://www.merckmanuals.com/en-ca/professional/endocrine-and-metabolic-disorders/diabetes-mellitus-and-disorders-of-carbohydrate-metabolism/diabetes-mellitus-dm

15. Rewers, M., & Ludvigsson, J. (2016). Environmental risk factors for type 1 diabetes. *Lancet (London, England)*, *387*(10035), 2340–2348. https://doi.org/10.1016/S0140-6736(16)30507-4

16. Pulgaron, E. R., & Delamater, A. M. (2014). Obesity and type 2 diabetes in children: epidemiology and treatment. *Current diabetes reports*, *14*(8), 508. https://doi.org/10.1007/s11892-014-0508-y

17. Ye, J. (2013). Mechanisms of insulin resistance in obesity. *Frontiers of Medicine, 7*(1), 14–24. https://doi.org/10.1007/s11684-013-0262-6

18. Punthakee, Z., Goldenberg, R., & Pamela, K. (2018). Definition, Classification and Diagnosis of Diabetes, Prediabetes and Metabolic Syndrome. *Canadian Journal of Diabetes*, *42*, S10-S15. https://doi.org/10.1016/j.jcjd.2017.10.003

19. Prebtani, A. P. H., Bajaj, H. S., Goldenberg, R., & Mullan, Y. (2018). Reducing the Risk of Developing Diabetes. *Canadian Journal of Diabetes*, *42*, S20–S26. https://doi.org/10.1016/j.jcjd.2017.10.033

20. Imran, S. A., Agarwal, G., Bajaj, H. S., & Ross, S. (2018). Targets for Glycemic Control. *Canadian Journal of Diabetes*, *42*, S42–S46. https://doi.org/10.1016/j.jcjd.2017.10.030

21. Evans PL, McMillin SL, Weyrauch LA, Witczak CA. (2019). Regulation of Skeletal Muscle Glucose Transport and Glucose Metabolism by Exercise Training. *Nutrients*. *11*(10):2432. doi: 10.3390/nu11102432. PMID: 31614762; PMCID: PMC6835691.

22. Goodyear LJ, Kahn BB. (1998). Exercise, glucose transport, and insulin sensitivity. *Annu Rev Med*. *49*:235-61. doi: 10.1146/annurev.med.49.1.235. PMID: 9509261.

23. Apovian, C. M., Okemah, J., & O'Neil, P. M. (2019). Body Weight Considerations in the Management of Type 2 Diabetes. *Advances in therapy*, *36*, 44–58. https://doi.org/10.1007/s12325-018-0824-8

24. Nathan, D., Davidson, M., DeFronzo, R., Heine, R., Henry, R., Pratley, R., & Zinman, B. (2007). Impaired Fasting Glucose and Impaired Glucose Tolerance: Implications for care. *Diabetes Care*, *30*(3), 753-759. doi: 10.2337/dc07-9920

25. Lipscombe, L., Booth, G., Butalia, S., Dasgupta, K., Eurich, D. T., Goldenberg, R., Khan, N., MacCallum, L., Shah, B. R., & Simpson, S. (2018). Pharmacologic Glycemic Management of Type 2 Diabetes in Adults. *Canadian Journal of Diabetes*, *42*, S88–S103. https://doi.org/10.1016/j.jcjd.2017.10.034

26. Herman, W. H., Ye, W., Griffin, S. J., Simmons, R. K., Davies, M. J., Khunti, K., Rutten, G. E., Sandbaek, A., Lauritzen, T., Borch-Johnsen, K., Brown, M. B., & Wareham, N. J. (2015). Early Detection and Treatment of Type 2 Diabetes Reduce Cardiovascular Morbidity and Mortality: A Simulation of the Results of the Anglo-Danish-Dutch Study of Intensive Treatment in People With Screen-Detected Diabetes in Primary Care (ADDITION-Europe). *Diabetes care*, *38*(8), 1449–1455. https://doi.org/10.2337/dc14-2459

27. Unnikrishnan, R., Shah, V. N., & Mohan, V. (2016). Challenges in diagnosis and management of diabetes in the young. *Clinical diabetes and endocrinology*, *2*(18). https://doi.org/10.1186/s40842-016-0036-6

28. Ganz, M.L., Wintfeld, N., Li, Q. *et al.* (2014). The association of body mass index with the risk of type 2 diabetes: a case–control study nested in an electronic health records system in the United States. *Diabetol Metab Syndr*, 6, 50. https://doi.org/10.1186/1758-5996-6-50

29. Ali O. (2013). Genetics of type 2 diabetes. *World journal of diabetes*, *4*(4), 114–123. https://doi.org/10.4239/wjd.v4.i4.114

30. Strachan, D., & Rose, G. (1991). Strategies of prevention revisited: Effects of imprecise measurement of risk factors on the evaluation of “high-risk” and “population-based” approaches to prevention of cardiovascular disease. *Journal Of Clinical Epidemiology*, *44*(11), 1187-1196. doi: 10.1016/0895-4356(91)90151-x

31. Craig, J. (2008) Complex diseases: Research and applications. *Nature Education* 1(1):184

32. McCulloch, D.W. (2019) Classification of diabetes mellitus and genetic diabetic syndromes. In: Mulder, J.E., Nathan, D.M., and Wolfsdorf, J.I. (Eds.), *UpToDate*. Retrieved March 17, 2021, from https://www.uptodate.com/contents/classification-of-diabetes-mellitus-and-genetic-diabetic-syndromes

33. Schork, N. J., Murray, S. S., Frazer, K. A., & Topol, E. J. (2009). Common vs. rare allele hypotheses for complex diseases. *Current opinion in genetics & development*, *19*(3), 212–219. https://doi.org/10.1016/j.gde.2009.04.010

34. Bermingham, M. (2017). Genetics of Type 1 Diabetes [Conference presentation]. Centre for Genomic and Experimental Medicine, Science Insights.

35. Ségurel, L., Austerlitz, F., Toupance, B., Gautier, M., Kelley, J. L., Pasquet, P., et al., (2013). Positive selection of protective variants for type 2 diabetes from the Neolithic onward: a case study in Central Asia. *European Journal of Human Genetics*, 21(10), 1146–1151. doi: http://doi.org/10.1038/ejhg.2012.295

36. Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., *et al*. (2009). Finding the missing heritability of complex diseases. *Nature*, *461*(7265), 747–753. https://doi.org/10.1038/nature08494

37. Schwarze, K., Buchanan, J., Taylor, J., & Wordsworth, S. (2018). Are whole-exome and whole-genome sequencing approaches cost-effective? A systematic review of the literature. *Genetics In Medicine*, *20*(10), 1122-1130. doi: 10.1038/gim.2017.247

38. Gibbs, R. (2020). The Human Genome Project changed everything. *Nature Reviews Genetics*, *21*(10), 575-576. doi: 10.1038/s41576-020-0275-3

39. LaFramboise T. (2009). Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic acids research*, *37*(13), 4181–4193. https://doi.org/10.1093/nar/gkp552

40. Loos, R. (2020). 15 years of genome-wide association studies and no signs of slowing down. *Nature Communications*, *11*(1). doi: 10.1038/s41467-020-19653-5

41. Vujkovic, M., Keaton, J., Lynch, J., Miller, D., Zhou, J., & Tcheandjieu, C. et al. (2020). aa *Nature Genetics*, *52*(7), 680-691. doi: 10.1038/s41588-020-0637-y

42. Läll, K., Mägi, R., Morris, A. et al. (2017). Personalized risk prediction for type 2 diabetes: the potential of genetic risk scores. *Genet Med* 19, 322–329 https://doi.org/10.1038/gim.2016.103

43. Lee, S., Abecasis, G., Boehnke, M., & Lin, X. (2014). Rare-Variant Association Analysis: Study Designs and Statistical Tests. *The American Journal Of Human Genetics*, *95*(1), 5-23. doi: 10.1016/j.ajhg.2014.06.009

44. Katsonis, P., Koire, A., Wilson, S. J., Hsu, T. K., Lua, R. C., Wilkins, A. D., & Lichtarge, O. (2014). Single nucleotide variations: biological impact and theoretical interpretation. *Protein science : a publication of the Protein Society*, *23*(12), 1650–1666. https://doi.org/10.1002/pro.2552

45. Zhang, X., Basile, A., Pendergrass, S., & Ritchie, M. (2019). Real world scenarios in rare variant association analysis: the impact of imbalance and sample size on the power in silico. *BMC Bioinformatics*, *20*(46). doi: 10.1186/s12859-018-2591-6

46. Guo, M. H., Plummer, L., Chan, Y. M., Hirschhorn, J. N., & Lippincott, M. F. (2018). Burden Testing of Rare Variants Identified through Exome Sequencing via Publicly Available Control Data. *American journal of human genetics*, *103*(4), 522–534. https://doi.org/10.1016/j.ajhg.2018.08.016

47. Flannick, J., Mercader, J.M., Fuchsberger, C. *et al.* (2019). Exome sequencing of 20,791 cases of type 2 diabetes and 24,440 controls. *Nature* 570, 71–76. https://doi.org/10.1038/s41586-019-1231-2

48. Flannick, J., Florez, J. (2016). Type 2 diabetes: genetic data sharing to advance complex disease research. *Nat Rev Genet* 17, 535–549. https://doi.org/10.1038/nrg.2016.56

49. Chheda, H., Palta, P., Pirinen, M. *et al.* (2017). Whole-genome view of the consequences of a population bottleneck using 2926 genome sequences from Finland and United Kingdom. *Eur J Hum Genet* 25, 477–484. https://doi.org/10.1038/ejhg.2016.205

50. Popejoy, A. B., Ritter, D. I., Crooks, K., Currey, E., Fullerton, S. M., Hindorff, L. A., Koenig, B., Ramos, E. M., Sorokin, E. P., Wand, H., Wright, M. W., Zou, J., Gignoux, C. R., Bonham, V. L., Plon, S. E., Bustamante, C. D., & Clinical Genome Resource (ClinGen) Ancestry and Diversity Working Group (ADWG) (2018). The clinical imperative for inclusivity: Race, ethnicity, and ancestry (REA) in genomics. *Human mutation*, *39*(11), 1713–1720. https://doi.org/10.1002/humu.23644

51. GenomeAsia100K Consortium., Wall, J.D., Stawiski, E.W. *et al.* (2019). The GenomeAsia 100K Project enables genetic discoveries across Asia. *Nature* 576, 106–111. https://doi.org/10.1038/s41586-019-1793-z

52. Huang, T., Shu, Y., & Cai, Y. D. (2015). Genetic differences among ethnic groups. *BMC genomics, 16,* 1093. https://doi.org/10.1186/s12864-015-2328-0

53. Ye, J. (2013). Mechanisms of insulin resistance in obesity. *Frontiers of Medicine, 7*(1), 14–24. https://doi.org/10.1007/s11684-013-0262-6

54. Guo, M. H., Plummer, L., Chan, Y. M., Hirschhorn, J. N., & Lippincott, M. F. (2018). Burden Testing of Rare Variants Identified through Exome Sequencing via Publicly Available Control Data. *American journal of human genetics*, *103*(4), 522–534. https://doi.org/10.1016/j.ajhg.2018.08.016

55. Francioli, L., Tiao, G., Karczewski, K., Solomonson, M., & Watts, N. (2018). gnomAD v2.1. *Broad Institute*. Retrieved August 22 2020, from https://gnomad.broadinstitute.org/blog/2018-10-gnomad-v2-1/

56. Monkol, L., Karczewski, K.J, Exome Aggregation Consortium, et. al (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536: 285-291. doi:10.1038/nature19057.

57. Naylor R, Knight J.A, del Gaudio D. (2018), Maturity-Onset Diabetes of the Young Overview. (Adam MP, Ardinger HH, Pagon RA, et al., editors). *GeneReviews* [Internet]. https://www.ncbi.nlm.nih.gov/books/NBK500456/

58. Barba, M., Czosnek, H., & Hadidi, A. (2014). Historical perspective, development and applications of next-generation sequencing in plant virology. *Viruses*, *6*(1), 106–136. https://doi.org/10.3390/v6010106

59. Scott, R. A., Scott, L. J., Mägi, R., Marullo, L., *et al*. (2017). An Expanded Genome-Wide Association Study of Type 2 Diabetes in Europeans. Diabetes, 66(11), 2888 LP – 2902. https://doi.org/10.2337/db16-1253

60. Gerstein, H.C. & Werstuck, GH. 2013. Dysglycaemia, vasculopenia, and the chronic consequences of diabetes. *Lancet Diabetes Endocrinol*. Sep;1(1):71-8. doi: 10.1016/S2213-8587(13)70025-1.

61. Burchum, J., and Rosenthal., L. (2019). *Pharmacology for Nursing Care, 10th Edition*. Elsevier.

62. Wessel, J., Majarian, T. D., Highland, H. M., Raghavan, S., Szeto, M. D., Hasbani, N. R., de Vries, P. S., Brody, J. A., Sarnowski, C., DiCorpo, D., Yin, X.,

Hidalgo, B., Guo, X., Perry, J., O'Connell, J. R., Lent, S., Montasser, M. E., Cade, B. E., Jain, D., … Manning, A. K. (2020). Rare Non-coding Variation Identified by Large Scale Whole Genome Sequencing Reveals Unexplained Heritability of Type 2 Diabetes. *MedRxiv*, 2020.11.13.20221812. https://doi.org/10.1101/2020.11.13.20221812

63. Ahlqvist, E., Storm, P., Käräjämäki, A., Martinell, M., Dorkhan, M., Carlsson, A., Vikman, P., Prasad, R. B., Aly, D. M., Almgren, P., Wessman, Y., Shaat, N., Spégel, P., Mulder, H., Lindholm, E., Melander, O., Hansson, O., Malmqvist, U., Lernmark, Å., … Groop, L. (2018). Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. *The Lancet Diabetes & Endocrinology*, *6*(5), 361–369. https://doi.org/10.1016/S2213-8587(18)30051-2

CHAPTER 2 – Methodology

2.1 Overview

This chapter covers thesis methodology in four sections. First, the general exome sequencing process is summarized. Then selection criteria used for Type 2 Diabetes (T2D) are described. Finally, the intracohort and inter-cohort bioinformatic analyses of exome sequenced variants with T2D are explained in separate sections.

2.2 Exome Sequencing

2.2.1 Introduction

There are currently several types of exome sequencing technologies, each patented by competing companies. There is much overlap between these, so a generalized overview of the exome sequencing pipeline is presented. However, important differences in exome sequencing workflows are highlighted when they cause bioinformatic implications downstream. To provide contrast, the popular Ion Torrent and Illumina platforms are primarily used as examples.

2.2.2 DNA Preparation

The exome sequencing workflow begins with collection of blood samples, followed by the extraction and purification of genomic deoxyribonucleic acid (DNA). The purified genomic DNA is then set to a specific input quantity, approximately in the range of 100 nanograms. Next, two comprehensive steps must be undertaken to qualify genomic DNA for exome sequencing: exome capture and library preparation. Exome capture is the isolation of protein coding regions, or exons, from the rest of the genome. Library preparation refers to the attachment of DNA fragments to synthetic oligonucleotides required for sequencing. There are a multitude of function-specific oligonucleotides, including sample-identifying barcodes, polymerase chain reaction (PCR) primers, and adaptors for substrate hybridization [1]. After these steps, the DNA typically undergoes a form of PCR-based amplification, then normalization to picomolar range, and finally are loaded onto chips in the sequencing platform.

Exome capture and library preparation are universal steps in exome sequencing. However, their respective methodology varies considerably between platforms. For example, the Ion Torrent pipeline first does exome capture, utilizing multiplex PCR on unamplified genomic DNA. This is followed by library preparation involving adapter ligation, which mediate attachment to specialized Ion Sphere Particles for enrichment [1]. The Illumina pipeline instead starts with library preparation, which involves an amplification of genomic DNA. Exome capture is achieved via hybridization of target sequences with oligonucleotide probes, which are in turn bound to microbeads for purification and enrichment [1]. The differences between these two methodologies comes into play downstream, during the treatment of the sequencing data. The pre-capture genomic amplification in the Illumina pipeline can result in PCR duplicates of the exons. These PCR duplicates are essentially artifacts that can cause an over-representation error, necessitating their removal. In contrast, generation of PCR duplicates during the Ion Torrent pipeline are intentional. In this scenario, removing them would cause an under-representation error. In addition to exome capture and library preparation, the specific sequencing methodology also has an influence on the data produced.

2.2.3 Exome Sequencing

The exomic DNA that are loaded onto chips serve as templates for which complementary strands are constructed. Sequencing methodologies differ by platform, but generally involve recording the incorporation of bases to the complementary strand during PCR . For example, the Ion Torrent S5 platform automates the supply of deoxynucleotide triphosphates (dNTPs) on the sequencing chip: adenosine triphosphate, guanosine triphosphate, cytosine triphosphate, and tyrosine triphosphate. The dNTPs are supplied one type at a time in cycles. Each time a base is incorporated by the polymerase, a hydrogen is released into the local solution. A semiconductor sensor detects the resulting pH change, which translates to a signal proportional to the number of bases incorporated during the nucleotide cycle [2]. As the complementary strands are physically completed, these signals correspond to base calls and are used to construct digital sequences. These are also known as sequencing reads.

Reads can be influenced by the sequencing methodology used to construct them. For example, the detection sensitivity of homopolymer regions – multiple adjacent bases of the same type – varies considerably between platforms. With semi-conductor sequencing, the quality of base calls is inversely proportional to homopolymer length due to signal saturation [3]. Two adjacent adenosine bases would produce a signal approximately double in intensity compared to a single adenosine. In contrast, the difference in signal intensity between seven and eight adjacent adenosines would be harder to discern, possibly causing a miscall. A miscalled homopolymer, such as 8 cytosines miscalled as 6 cytosines, will lead to gaps in the alignment. Other platforms are more resilient to miscalling homopolymers, such as Illumina's iSeq platforms where fluorescently tagged nucleotides have a terminator that prevent incorporation of multiple nucleotide per cycle [2]. While homopolymer miscalls can be addressed by the alignment software, the process is specific to the sequencing platform used to generate the data.

2.2.4 Read Alignment and Quality Control

Since the exomic DNA is amplified during library preparation, the raw sequencing reads consist of multiple, overlapping strands (Figure 2.1a). Prior to alignment, reads are trimmed of adapter sequences and low quality 3' ends, which are sequenced last. Furthermore, trimmed reads are excluded if: they are too short in base length, contain adapter dimers, lack barcode sequences, or are polyclonal in origin. The trimmed reads that pass these quality checks are then be aligned to a human reference genome assembly (Figure 2.1b). Alignment software generates candidate mapping locations and use an alignment algorithm, such as Smith Waterman, to create multiple alignment sets. The sets are aggregated and those with the highest overall mapping quality are saved as Binary Alignment Map (BAM) files.

Samples, represented in the BAM files, are assessed on several metrics. Many utilize the read depth (DP) value, which is the number of reads that contribute to a given base (Figure 2.1c). The distribution of DP across the exome, or coverage, is not uniform and can vary between samples and different next generation sequencing platforms [4]. Three DP-based metrics are the mean DP, proportion covered by a minimum number of reads, and the proportion covered within a percentage of the mean DP. Other metrices can be the proportion of reads mapped to specific target regions and proportion of bases with a minimum Phred-based quality score. Samples that fail to satisfy these metrics can be salvaged by repeating in their library preparation and sequencing, then comparing to samples of high-quality metrics as references.[4]. The more samples there are in an exome sequencing study, the better the ability to detect high- and low-quality samples.

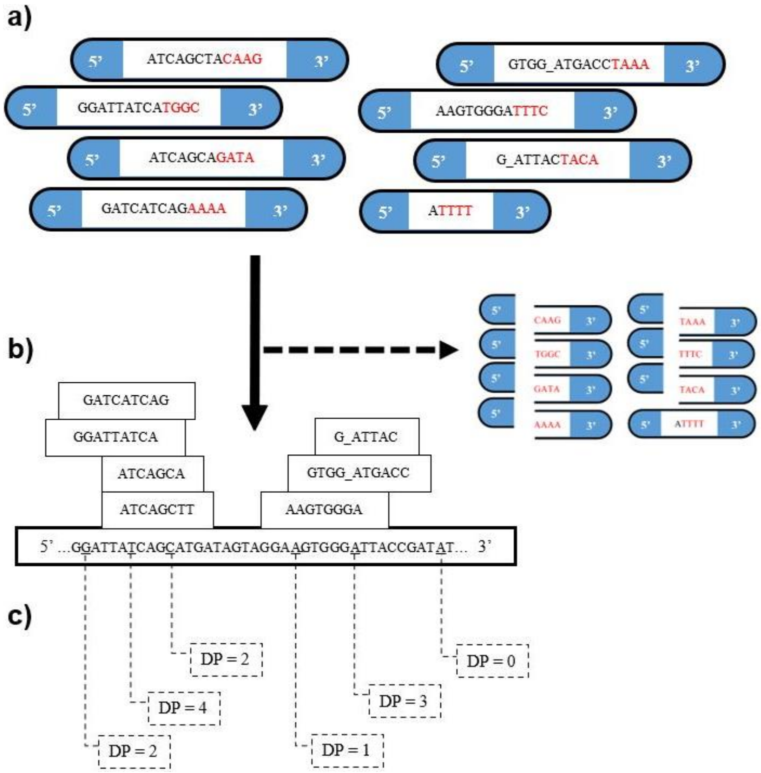


Figure 2.1: Sequencing read quality control and alignment, a) Raw sequencing reads are flanked by adaptor sequences (blue) and 3' bases are low quality, b) Quality control includes trimming of adaptor sequences, low quality 3' bases, and reads that are too short in length. The remaining reads are then aligned to sequences from the reference genome, c) Read depth for a given base determined by the number of overlapping reads containing that base.

2.2.5 Variant Calling

Finally, the sample BAM files are used to call variants, which are alleles at any given genomic location that are different from the reference. For example, if the reference allele is an adenosine, and the sample allele is not adenosine, an alternate variant is called at that location. Single nucleotide variants (SNV) are the most common kind of variant, where a single nucleotide base has been substituted for another. When a nucleotide base is gained or lost relative to the reference genome, insertion or deletion (Indel) variants are called, respectfully. Examples of variant types and possible mutations are summarized in Table 2.1.

Table 2.1: Variant types and possible mutations

Variant type	**Example alleles (Reference, Alternate)**	**Example protein-altering mutations**
SNV	Adenosine, Cytosine	Non-synonymous
SNV	Adenosine, Tyrosine	Stop-gain or stop-loss
SNV	Adenosine, Guanine	Splicing
Insertion indel	Adenosine, Cytosine-Guanine	Frameshift
Deletion indel	Adenosine, nothing	Splicing
Deletion indel	Adenosine-Guanine-Guanine, nothing	Non-frameshift

Each exome sequencing platform comes with its own software suite that handles the data processes described thus far, as well as variant calling. While variant calling specifications and capabilities differ between the software suites, they all start with individual BAM files and end up with the variant calls of all samples consolidated into a single table. Briefly, reads from individual samples are first compared against the reference genome, identifying the alternate variants. At a given genomic location, the proportion of reads with the alternate variant is used to statistically infer the genotype: heterozygous alternate genotype (0/1) or homozygous (1/1) alternate. If the reads at a genomic location do not consistently have the same variant, or if they are of otherwise low quality, a non-PASS tag is assigned for later exclusion. Then the samples are compared against each other. Variants with the homozygous reference genotype (0/0) are identified if alternate variant calls are made in some samples but not others. Variants with homozygous reference genotypes in all samples are excluded; their genotype information undistinguishable from the reference genome. The missing genotype (./.) can be called for a variant in some samples if read constancy or sample-wide representation is low. Excluding low quality variants is necessary because they may be artifacts and therefore represent false positives. All the sample genotypes are translated into a numerical code and combined into the table, often arranged in the variant call format (VCF). Each row of the VCF represents a different variant, with the columns designated to information about each variant: its genomic coordinates (chromosome, position, database identification number, reference and alternate allele), then various metrics like DP or Phred quality scores, and finally the genotypes of each individual. The tabular VCF, as well as alternative binary formats, serve as accessible jumping off points for bioinformatic analysis of exome sequencing data.

2.3 T2D Selection Criteria

Exome sequences of T2D cases and controls were obtained from the United Kingdom Biobank (UKB) and studies from the Database of Genotypes and Phenotypes (DBGap). 200,643 exomes were downloaded from the UKB successfully. However, the ongoing coronavirus 2019 pandemic led to a delay in DBGap data availability. Of the 26 studies requested, only 8 were approved for access, and of those only 3 were downloaded: the Korea Association Research (KARE) project (n=1087), the Metabolic Syndrome in Men Study (METSIM) (n=982), and the San Antonio Mexican American Family Studies (SAMAFS) (n=491).

Individuals in the DBGap cohorts had assigned phenotype for T2D case or control status. Inclusion criteria were not identical across all studies, but they shared usual qualifiers such as past T2D diagnoses, sufficient fasting or 2-hour plasma glucose levels, and family history [19,20,21,37]. In the UKB, T2D status was not preassigned and had to be determined using provided International Classification of Diseases, Tenth Revision (ICD-10) codes. Guidelines on T2D case and control selection from Choi, *et al.* (2019) and Khera, *et al.* (2018) were also consulted. [32, 33]. Table 2.2 summaries T2D case and control selection criteria for the UKB, KARE, METSIM, and SAMAFS. Since it only offers summary level information, T2D controls could not be individually selected from GnomAD. Curation of the database should have prevented enrichment of T2D cases relative to the general population, though phenotypic heterogeneity is still possible because there were no exclusion criteria specific to diabetes.

Table 2.2: T2D case and control selection criteria

Cohort	**Case criteria**		**Control criteria**	
UKB	ICD-10s present:	First occurrence, underlying / contributory cause of death:	ICD-10s absent:	First occurrence, or underlying / contributory cause of death:
	E11	Non-insulin-dependent diabetes mellitus, ...	E10	Insulin-dependent diabetes mellitus
	E11.0	with coma	E11	Non-insulin-dependent diabetes mellitus
	E11.1	with ketoacidosis	E12	Malnutrition-related diabetes mellitus
	E11.2	with renal complications	E13	Other specified diabetes mellitus
	E11.3	with ophthalmic complications	E14	Unspecified diabetes mellitus
	E11.4	with neurological complications	E15	Nondiabetic hypoglycaemic coma
	E11.5	with peripheral circulatory complications	E16	Other disorders of pancreatic internal secretion
	E11.6	with other specified complications		
	E11.7	with multiple complications		
	E11.8	with unspecified complications		
	E11.9	without complications		
KARE	Past history of T2D.		No past history of diabetes.	
	Use of T2D medication.		No anti-diabetic medication.	
	Fasting plasma glucose >/=7 mmol/l or plasma glucose >/=11.1 mmol/l 2 hours after ingestion of 75gm oral glucose load.		Fasting plasma glucose <5.6 mmol/l and plasma glucose 2 hours after ingestion of 75g oral glucose load <7.8 mmol/l at both baseline and follow up timepoints.	
	Age of disease onset >/=40 years.		Older subjects with normal glucose prioritized.	
	Participants with early onset and family history prioritized.		Samples with age of diagnosis <40 excluded.	

SAMAFS	Unrelated samples. Fasting plasma glucose of 7.0 mmol/l or a 2-h glucose after an oral glucose tolerance test of 11.1 mmol/l Self-reported physician-diagnosed diabetes. Reported current therapy with either oral antidiabetic agents or insulin. Family history of diabetes.	Unrelated samples.
METSIM	Previous diagnosis of T2D, or both fasting and 2-hr criteria met for new T2D diagnosis Family history of diabetes. Anti-GAD antibody <50 U/mL to rule out Type 1 Diabetes. C-peptide >0.10 nmol/L. Unrelated samples. Preferentially select individuals with genotype data, as well as non-genotyped individuals with earlier possible age of diagnosis.	Normal glucose tolerance at baseline and follow-up visits. Prioritized samples with no family history of diabetes and meeting strict normal glucose tolerance criteria: fasting glucose <5.6 mmol/l and 2 hour post-challenge glucose <7.8 mmol/l. Additional samples selected with fasting glucose <6.1 mmol/l and 2 hour post-challenge glucose <7.8 mmol/l. Unrelated samples. Older controls preferentially selected.

2.4 Intracohort Analyses

2.4.1 Introduction

The ideal goal of this research would be to identify protein altering rare variants (RVs) which have effects on T2D risk, as defined by the selection criteria in the previous section. However, individual RVs are unlikely to occur in many samples due to their low population frequency, limiting power to detect association signals [16,18]. Instead, the RV gene burden method was employed, where in each gene the effects of RVs were cumulated together into RV gene burdens [18]. RV gene burden were assumed to have loss of function (LOF) effects on the encoded proteins [18]. Any gene casual in T2D pathogenesis would be expected to have a significantly greater frequency or magnitude of RV gene burden among carriers in cases than in controls [18]. Therefore, the actual goal of this research was to discover genes that, through their RV gene burden, contribute to T2D risk.

The ability of RV gene burden to predict T2D case or control status was assessed with logistical regression in intracohort analyses of the UKB and repeated in KARE, METSIM, SAMAFS, as well as a pooled meta-analysis of the latter three cohorts. However, between downloading public exome sequencing data and RV gene burden analyses several quality control, annotation, and filtering steps were required. The numbers of samples, genes, and variants in each step, across all cohorts, are summarized in Table 2.3. A schematic overview is shown in Figure 2.2.

Table 2.3: Intracohort sample, variant, and gene counts

Study	Initial download (samples / variants)	Quality control (samples / variants)	Rare protein altering variants	T2D cases	Non-T2D controls*	Number of Genes
UKB	200,643 / 17,975,236	173,688 / 17,570,704	2,994,778	8,784	153,431	18,815
KARE	1087 / 690,291	973 / 631,648	116,119	462	511	11,927
METSIM	982 / 478,498	969 / 456,299	78,772	477	492	9,606
SAMAFS (Project 1)	491 / 549,096	309 / 431,740	51,277	168	141	7,960

11,473 samples were excluded from non-T2D control status in the UKB because they belonged to confounding disease groups, such as Type 1 Diabetes. SAMAFS project 2 was not used because it was entirely comprised of related individuals.

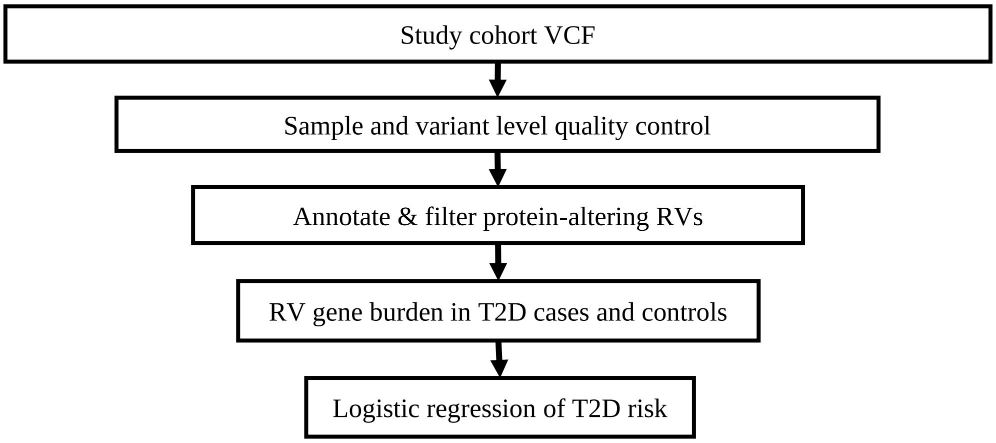

Study cohort VCF
Sample and variant level quality control
Annotate & filter protein-altering RVs
RV gene burden in T2D cases and controls
Logistic regression of T2D risk

Figure 2.2: Intra-cohort analyses overview. Flowchart summarizing the steps involved in the intra-cohort analyses of T2D cases and controls.

2.4.2 Additional Quality Control

Individual-level exome sequencing data for KARE, METSIM and SAMAFS was downloaded from DBGap as VCFs. The UKB's exome sequence data release in joint call set PLINK format was downloaded from the UKB data showcase [9]. Following download, additional sample and variant level quality control was performed in all cohorts. The quality control steps done in the UKB are shown as an example in Figure 2.3.

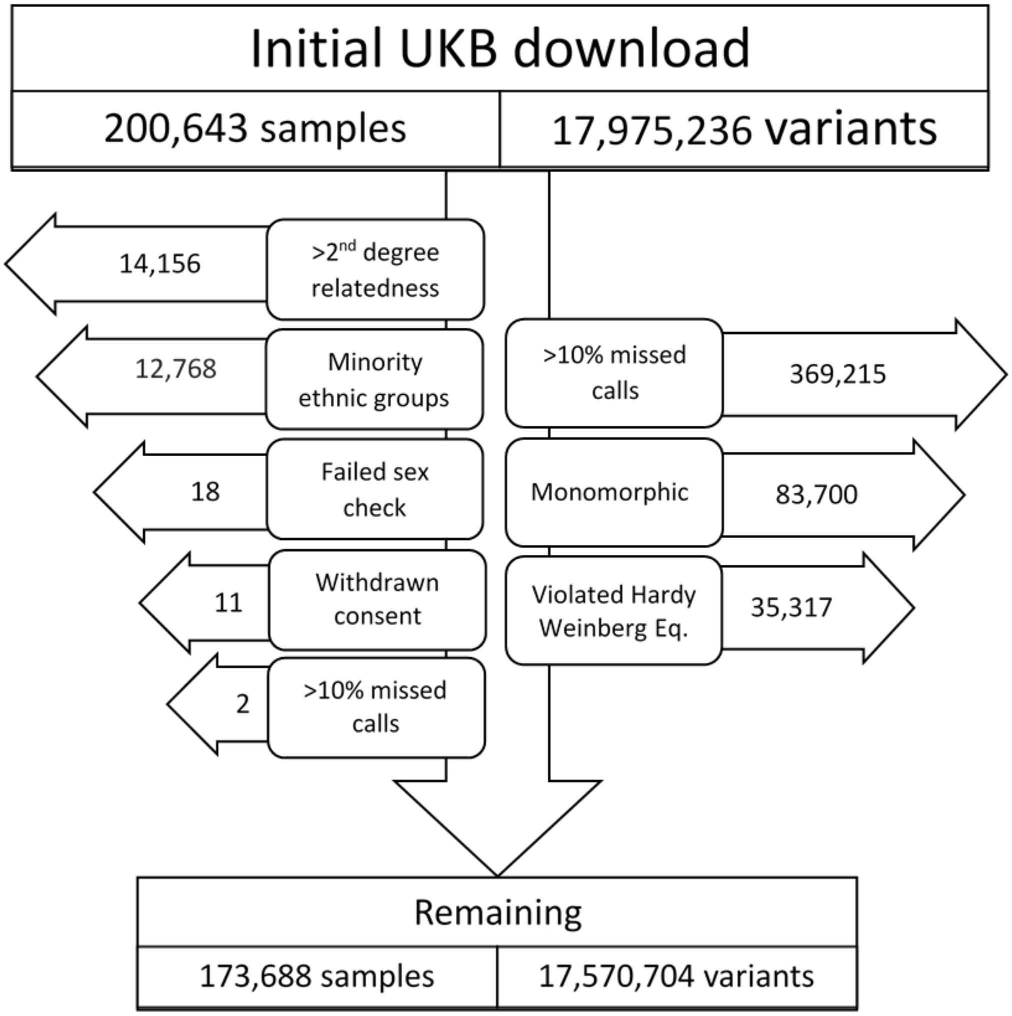


Figure 2.3: Flowchart of quality control for UKB exome sequencing data

Samples were excluded if they failed to meet criteria in five factors: non-majority ethnicity and ancestry, genetic relatedness, proportion of missing variant calls, sex checks, and current consent status (Figure 2.3). Only samples belonging to each cohort's majority ethnicity or ancestry were included: British in the UKB, Korean in KARE, Finnish in METSIM, and Latino in SAMAFS. Samples were excluded if they had greater than a 2nd degree of genetic relatedness with other participants. The familial inheritance of alleles in related samples would confound T2D allele burden in the majority unrelated samples, necessitating removal of the former [5,22]. The threshold of missing variant calls was set to a maximum proportion of 10% to ensure a high quality of variant calls. Samples were excluded if genetic sex, as determined by markers on the X chromosome, conflicted with reported gender [6]. While alternative biological explanations, such as hormone insensitivity, could explain a failed sex check, experimental or human error are assumed [6,9]. Missingness and sex checks were done using PLINK's *missing* and *check sex* functions, respectfully [20]. Relatedness was evaluated with KING software [21]. Finally, samples were removed if consent for research use was withdrawn [7].

Variants were excluded based on four factors: monomorphism, high degree of missing variant calls, lack of adherence to Hardy Weinberg Equilibrium (HWE), and location in under-called regions (Figure 2.3) [25,26]. Monomorphic variants have minor allele frequencies (MAFs) of 1 or 0, and therefore cannot be analyzed [8]. Like with the samples, variants with a proportion of missed calls greater than 10% were excluded. HWE describes the distribution of alleles in the absence of evolutionary forces and defined by the equation $q^2 + 2qp + p^2$, where p and q are recessive and dominant alleles [9,21-3]. Variants that violated

HWE below a set threshold of $p < 5 \times 10^{-6}$ were indicative of sequencing error. Following these steps, the 22 autosomal chromosomes were retained for analysis. The sex chromosomes were excluded because they require stratification by males and females, and they also present analytical challenges. For example, in females X-inactivation of the maternal or paternal chromosome is random in each cell and is difficult to identify with current sequencing technology [38].

2.4.3 Format Standardization

While the same level of quality was achieved across the various studies, further standardization was required. First all variants were put into a biallelic format, such that at any given site only two alleles were present: the reference and a single alternate. Multiallelic sites with more than two alternate alleles were split into separate biallelic variants because they need be assessed on an individual basis. Second, insertion and deletion (indel) variants were normalized. Normalization involves two criteria: left-alignment and parsimony. Left-alignment ensures that a variant's position is as far left as possible and parsimony allows only the shortest length of nucleotides to represent the variant [10]. These steps were done using *vcfbreakmulti* and *norm* functions of *vcf library* and *bcftools*, respectfully [27].

2.4.4 Variant Annotation

By this point the processed information for all cohorts was limited to variant lists, sequencing metrics and sample genotypes. Additional context was needed to conduct association analysis, namely: gene information, mutation outcomes, superpopulation

MAFs, and pathogenicity predictions. These were to be used to find protein altering mutations.

Gene and mutation annotations were done by cross-referencing variant coordinates with the RefGene database, accomplished using the *Annovar* software [11,29] . This added three essential pieces of information for each variant: the gene it resides in, whether it is in a coding region, and, if so, the alternate allele's effect on the amino acid. While exome capture is centered around exons, intronic regions would still have been sequenced. Reads capturing the 5' and 3' ends of exons need to overlap with the surrounding introns [1]. This extra non-exonic sequencing is not superfluous because splicing sites can be in introns [1]. Any alternate variant in an exon will change its codons, which has the potential to change the corresponding amino acids. Some mutations are so severe that not only protein alteration, but complete LOF in the peptide is all but guaranteed. A likely LOF scenario is where indels of a base length indivisible by three result in a frameshift mutation, upsetting the whole open reading frame. Conversely, the functional impact of SNVs are harder to define, since they can either be synonymous or nonsynonymous. The former, where the codon change still results in the same amino acid, is usually considered benign. Nonsynonymous SNVs will always code for a different amino acid, however the consequence can vary considerably. Possibilities range from a minor effect on tertiary structures or polarity to completely altering the resultant protein [12]. For example, cysteine-altering residues can disrupt disulfide bridge formation and cause massive destabilization. Nonsynonymous SNVs can also cause LOF scenarios by prematurely activating stop codons or deactivating start codons, in stopgain or startloss mutations, respectfully.

Deactivation of Stop codons in stoploss mutations will likely create an elongated protein and be LOF [13]. Estimating the effects of nonsynonymous SNVs can be further refined by considering their population frequency as a function of natural selection.

Variants were annotated with observed MAFs in GnomAD's five major superpopulations: non-Finish European (NFE), East Asian (EAS), South Asian (SAS), American admixed/Latino (AMR), African/African American (AFR). Additionally, METSIM variants were annotated with the minor Finnish (FIN) MAF. All cohorts were also annotated with an internal MAF specific to each, calculated using *vcftools freq* function [31]. A rare variant in any given single ancestry may not be solely explained by negative selective pressure. *De novo* (new) mutations and sequencing artifacts would present as rare. Genetic bottlenecking and genetic drift could alter the distribution of pathogenic variants, independent of impact on survival [14]. Regional recruitment biases, methodological biases, or other population homogeneities could likewise skew distributions [15]. Therefore, assessing the rarity across multiple ancestries helps detect variants that are truly deleterious.

The final annotation was the Mendelian Clinically Applicable Pathogenicity (MCAP) score. MCAP employs an ensemble of pre-existing scores and machine learning to predict the functional impact of nonsynonymous SNVs [15]. MCAP incorporates a variety of factors, including estimated impact on protein structure, evolutionary conservation rates, and models trained on verified pathogenic variants from Human Gene Mutation Database [15]. The MCAP score itself is a continuous numerical scale ranging from benign at 0 to pathogenic at 1.

2.4.5 Variant Filtering

The annotations were then used to select variants most likely to have a pathogenic influence: rare protein altering mutations. First, variants were included if they were located within exons or splicing regions. Then variants causing nonsynonymous SNV, frameshift insertion, frameshift deletion, stop-gain, or stop-loss mutations were selected. Of these, variants were excluded if they had a MAF greater than 0.01 in any of the five major GnomAD superpopulations and each cohort's internal frequency. METSIM variants were also compared to frequencies in GnomAD's minor FIN superpopulation. Finally, nonsynonymous SNVs were included if they had an MCAP greater than 0.025, which is the threshold recommended by the score's authors for optimal detection of true positives [15].

2.4.6 RV Gene Burden

To facilitate the discovery of genes with protein altering RVs that contribute to T2D risk, RV gene burdens were calculated. First, genotypes of individual RVs were converted to numeric allele burden scores using the additive penetrance model (Table 2.4). Under this model, it is assumed that each allele acts equally in magnitude and uniform in direction of effect [18]. This assumption allowed allele burden scores in each gene to be simply added together to form RV gene burdens. RV gene burden was calculated for all samples simultaneously.

Table 2.4: Genetic models of penetrance

Variant genotype	Genotype code	Additive allele burden	Dominant allele burden	Recessive allele burden
Missing	./.	0	0	0
Homozygous reference	0/0	0	0	0
Heterozygous alternate	0/1	1	1	0
Homozygous alternate	1/1	2	1	1

Allele burden in carriers of alternate mutations under different genetic models of penetrance. Allele burden is continuous in the additive model, while it is binary in the dominant and recessive models.

These calculations were merged with sample phenotypes, resulting in a large table for each cohort detailing each sample's age, sex, T2D case/control status, and the RV gene burdens of the different genes (Table 2.5). Underpowered genes with fewer than one carrier among cases or controls were removed from the analyses. The number of genes in remaining in each cohort after this pruning step are shown in Table 2.3. The UKB included additional fields for its top 40 principal components (PCs), which are sets of constructed values that explain variation in the cohort's genetic population structure [39]. Along with age and sex, the PCs were included in the logistic regression models used to predict the relationship of gene's RV burden with T2D.

Table 2.5: Example merged table of sample phenotypes, PCs, and RV gene burdens

Sample ID	Age	Sex	T2D	PC1	PC2	*GeneA*	*GeneB*
SAMPLE_001	37	0	0	-14.409	4.312	0	3
SAMPLE_002	45	0	1	-13.034	-2.191	0	0
SAMPLE_003	56	1	0	12.401	-4.093	0	1
SAMPLE_004	40	1	1	15.091	2.239	1	2

In this example, RV gene burden of four samples are shown in two fake genes, as well as age, sex, status as a T2D case or control, and the first two principal components (PC1 and

PC2). Since the additive allele burden model was used, RV gene burdens >= 3 are possible in samples if multiple variants are detected in a gene.

2.4.7 Penetrance and Logistic Regression

The effect RV gene burden had on T2D was presented in two ways: penetrance estimates and predictive logistic regression models.

In the context of this thesis, genetic penetrance can be defined as the probability that carriers, or samples with non-zero RV gene burdens, present as T2D cases [40]. The calculation of a gene's penetrance is straightforward:

$$Penentrance\ = \frac{number\ of\ carriers\ that\ are\ cases}{total\ number\ of\ carriers}$$

The ability of RV gene burden in each gene to predict T2D risk was assessed using logistic regression and adjusted for age and sex in all cohorts, as well as the 40 PCs in the UKB:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_{burden}X_{burden} + \beta_{age}X_{age} + \beta_{sex}X_{sex} + \beta_{PC1}X_{PC1} + \beta_{PC2}X_{PC2} + \beta_{PC3}X_{PC3} + \ldots\ldots\ldots \beta_{PC40}X_{PC40}$$

Where the logistic regression β_{burden} coefficient is the expected change in log odds of having the outcome (T2D) per unit increase (number of alleles) in the exposure (RV gene burden). The odds ratio (OR) is the exponent of the β_{burden} coefficient:

$$OR = exp(\beta_{burden})$$

OR shows the change in odds for developing T2D per each additional allele in RV gene burden. Uncertainty in the β_{burden} coefficient was represented by 95% confidence intervals calculated from each gene's standard error using the *confint* function from the Modern Applied Statistics with S package [41].

2.4.8 Exome-Wide Significance

Each estimate of a gene's OR had an accompanying P-value, which was the probability of observing the result of the logistic regression if the null hypothesis was true. In the context this project:

Test = Whether a gene's RV gene burden (exposure) predicts T2D risk (outcome)
Result = The OR between exposure and outcome
Null hypothesis = The OR equals 1.0 (does not predict)

Statistical significance is usually defined when the p-value is less than 0.05, a maximum 5% chance the result is false positive. However, the more tests (RV gene burden in each gene) conducted on the same sample (of T2D cases and controls), the greater the chance of getting false positive results [16]. Bonferroni correction accounts for this by adjusting the p-value threshold for statistical significance, p_{min}, by *n* number of individual tests [16]:

$$p_{min} < \frac{0.05}{n}$$

Exome-wide significance p_{min} was configured using the largest single-sample number of tests among the analyses: 18,815 genes in the UKB:

$$p_{min} < \frac{0.05}{18815}$$

$$p_{min} < 2.65 \times 10^{-6}$$

However, this threshold could be too conservative because genes are not necessarily independent from one another [34]. In a phenomenon known as linkage disequilibrium (LD), it is possible that two or more alleles can have a non-random association. There are many causes of LD, including genetic drift, chromosomal proximity, or selective pressures [35]. For example, protein altering RVs in the same gene could be in LD because of shared functional impact. While RVs have been thought to be free of LD, this may have been due to the inability to properly detect the phenomenon with past data available [36]. The outright dismissal of genes below the exome wide significance threshold as false positives may miss some otherwise insightful findings. Therefore, genes with P-values below 0.001 were also examined as suggestive associations.

2.4.9 DBGap Meta-analysis

A pooled meta-analysis was conducted with RV gene burdens from the three DBGap studies: KARE, METSIM, and SAMAFS. This was achieved by simply combining the merged phenotype and RV gene burden tables from each study. Exclusion of underpowered genes was reassessed, with the minimum number of carriers now counted across the three studies. For the 12,744 remaining genes, logistic regression of the ability of RV gene burden to predict T2D risk proceeded with the expanded pool of samples.

2.5 Inter-cohort Analyses

2.5.1 Introduction

In the inter-cohort analyses, RV Gene burden was compared between T2D cases in the UKB and DBGap studies and controls represented by GnomAD. GnomAD version 2.0.1 summary-level exome sequence VCFs and coverage information was downloaded from the Broad Institute website [6]. Preparation of exome sequencing data of T2D cases and GnomAD was similar to the intracohort analyses, but there were some important distinctions. These were the intersection of high coverage regions, the different method used to calculate RV gene burden in GnomAD, the use of delta counts and Z-tests, and the pursuit of inter-cohort corrections. In GnomAD and each corresponding T2D case cohort, the numbers of genes and variants at each major step are summarized in Table 2.6. An overview of the inter-cohort analyses is shown in Figure 2.4.

Table 2.6: Inter-cohort variant and gene counts

Study	Downloaded GnomAD variants	Intersected variants (GnomAD / study)	Protein altering variants (GnomAD / study)	Number of genes
UKB	13,146,649	8,545,792 / 7,668,684	1,263,887 / 1,207,653	15,534
KARE		505,560 / 372,699	71,251 / 159,995	5,835
METSIM		326,436 / 262,118	38,801 / 85,033	4,927
SAMAFS		422,733 / 245,056	43,718 / 35,746	2,326

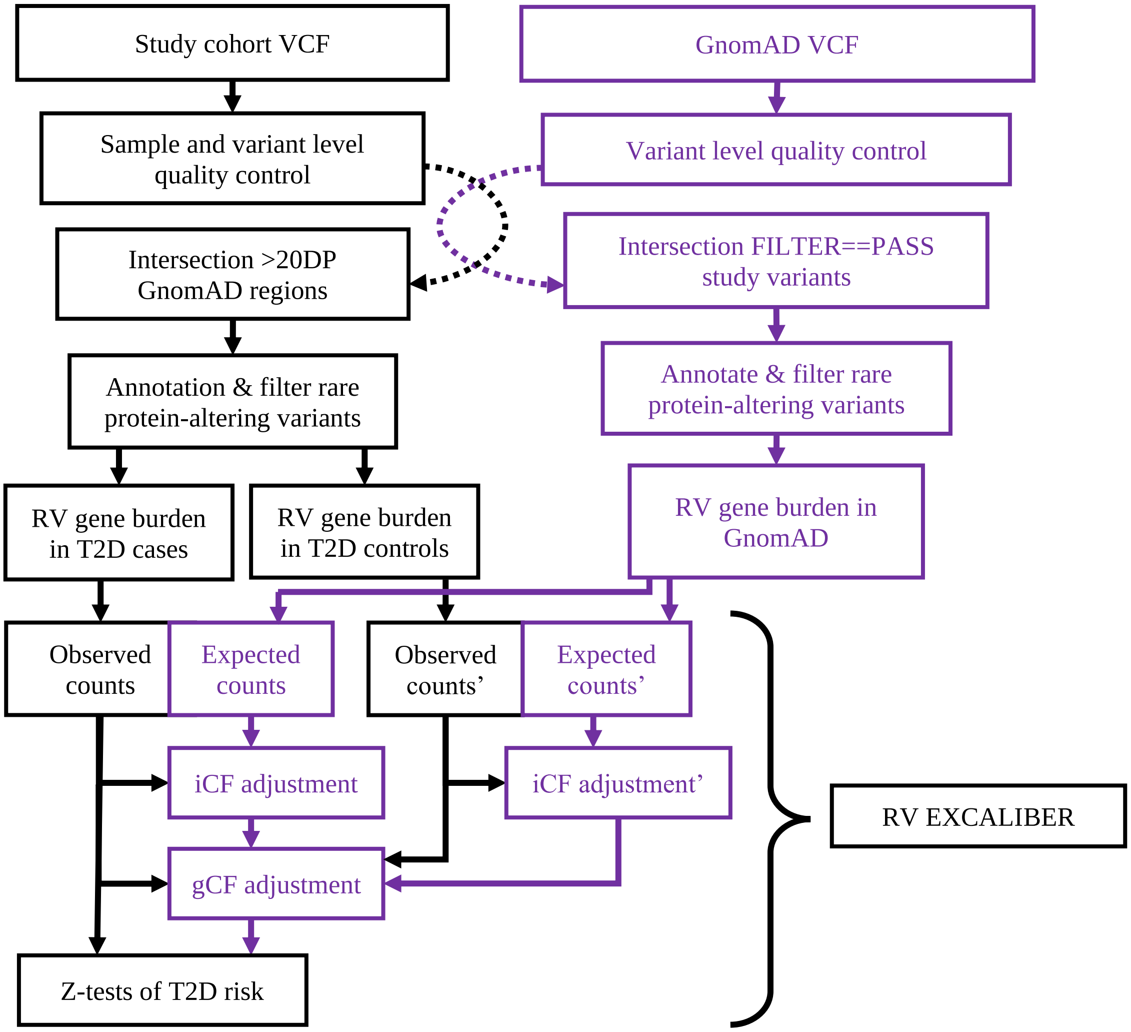

Study cohort VCF
GnomAD VCF
Sample and variant level quality control
Variant level quality control
Intersection >20DP GnomAD regions
Intersection FILTER==PASS study variants
Annotation & filter rare protein-altering variants
Annotate & filter rare protein-altering variants
RV gene burden in T2D cases
RV gene burden in T2D controls
RV gene burden in GnomAD
Observed counts
Expected counts
Observed counts’
Expected counts’
iCF adjustment
iCF adjustment’
gCF adjustment
Z-tests of T2D risk
RV EXCALIBER

Figure 2.4: Inter-cohort analyses overview. Flowchart summarizing the steps of the inter-cohort analyses of T2D cases and GnomAD.

2.5.2 Intersection

Non-differential coverage between GnomAD and each study was required. In this context, coverage refers to exomic regions that are successfully sequenced [17]. For a given base, its successful sequencing requires sufficient DP for a high-quality variant call. Coverage is therefore evaluated by the DP of regions across the exome. When detecting rare alternate alleles, regions with low DP are prone to produce false negatives.

Furthermore, coverage between cohorts can vary wildly because of differences in sequencing methodology and study population [18]. Avoidance of false negatives and positives can be achieved with intersection of high DP regions in both cohorts to achieve non-differential coverage.

Regional coverage depth can only be determined using the BAM files produced during the exome sequencing pipeline. BAM files contain DP information for the entirety of the aligned exon reads, allowing for a comprehensive evaluation of coverage. In contrast a VCF retains DP information only for the included variants. The large size of BAM files makes them restrictive to work with, necessitating consolidation of coverage information into the browser extensible data (BED) file format. BED files simply list the sequenced exonic regions, each denoted by a chromosome number and positional range from beginning to end. To establish regions of high coverage, BED files can be created that include regions that meet a minimum DP.

For the UKB and DBGap datasets, regional coverage could not be established. The UKB exome sequencing data release did include CRAM files (a compressed version of the BAM format), but the 175 terabytes of required disk space exceeded available storage. None of the exome sequencing data sourced from DBGap studies included coverage information. While GnomAD did not provide individual level BAM files, the database did include summary level coverage information in a BED-like format. From these, GnomAD regions with a minimum 20X DP were established. With the data available, two kinds of intersection were conducted.

In the first kind, variants from the UKB and DBGap study cohorts were independently intersected with the minimum 20X DP regions in GnomAD, resulting in variants that fell within those regions (Figure 2.5a). This was done with Bedtool's *intersectBed* function [24]. The UKB exome sequencing data was built using the Genome Reference Consortium Human genome build 38 (GRCh38) human genome reference assembly, whereas the GnomAD v2.0.1 had used the older GRCh37. To ensure compatibility, genomic coordinates were lifted over between the two assemblies using the University of California Santa Cruz liftOver executable and hg38 to hg19 chain file [11, 31]. While this achieved true non-differential coverage for UKB and the DBGap variants with respect to high coverage regions in GnomAD, intersection the opposite direction required compromise.

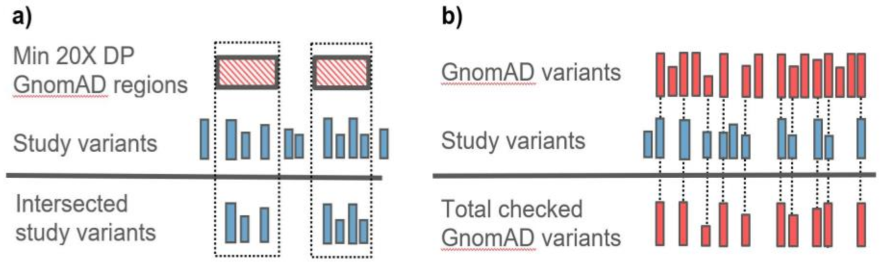

a)
Min 20X DP GnomAD regions
Study variants
Intersected study variants
b)
GnomAD variants
Study variants
Total checked GnomAD variants

Figure 2.5: Intersection of high coverage regions and variants in GnomAD and study cohorts, a) Variants from a study cohort are excluded if they fall outside of high coverage GnomAD regions, b) When there is insufficient coverage information, GnomAD variants are excluded if they do not share Pass sites with study variants.

Variants that qualified with the Pass filter on the GnomAD VCF were checked against UKB and DBGap variants that qualified with the Pass in their respective VCFs, resulting in *total checked* GnomAD variants (Figure 2.5b). The Pass filter is based on quality scores, which in turn are based on DP. Therefore, the total checked GnomAD variants approximate non-differential coverage with respect to high coverage regions in UKB and the DBGap cohorts. This is not a perfect substitution because GnomAD variants, which could otherwise exist within high coverage regions, will be excluded if coordinates are not an exact match. After intersection, the number of variants in GnomAD exceeded those in each study cohort, though this ratio equalized or reversed after filtering of protein altering RVs (Table 2.5). Since each cohort is much smaller compared to GnomAD, the intersection may have been too stringent and led to under-estimates of RV gene burden. Doing either no intersection or a one-way approach (Figure 2.5a) with the limited coverage information available may have been better options. Ultimately, the two-way intersection was used because it best emulated the original regional coverage method and GnomAD was considered to have enough variants to handle the added stringency.

2.5.3 GnomAD RV Gene Burden

RV gene burden in GnomAD was calculated using the MAF of the ancestry corresponding to each study cohort. For example, allele counts (AC) of UKB T2D cases were compared to the NFE MAF of variants in GnomAD. Per gene variant MAFs were added together to form cumulative minor allele frequencies (CMAFs). The GnomAD CMAFs were then multiplied by two which converted them into cumulative minor allele counts (CMACs) relative to the number of alleles in humans who are diploid. Corresponding to each study gene matrix was a GnomAD table that listed all genes and their CMACs. Since the CMACs were based on GnomAD's population MAFs, they could be scaled up by multiplying by the sample size of the corresponding study.

2.5.4 Delta Counts

Despite the earlier intersection of high coverage regions, a direct comparison of allele burden between study cohorts and GnomAD will still run into confounding issues. For example, it is impossible to determine from summary level data if the RVs were inherited together in groups, a phenomenon called linkage disequilibrium (LD) . Since LD leads to inflated allele burden, a case control comparison using Fisher's exact test or logistic regression would encounter many false positives [18].

A solution was to do delta counts for each sample of each gene by finding the difference in allele count between the observed T2D cases and expected controls (GnomAD). The null hypothesis for each gene was no difference, or a delta count of zero, which could be checked with a z-test:

$$Z = \frac{mean\ delta\ count}{\left(\frac{SD\ of\ delta\ count}{\sqrt{sample\ size}}\right)}$$

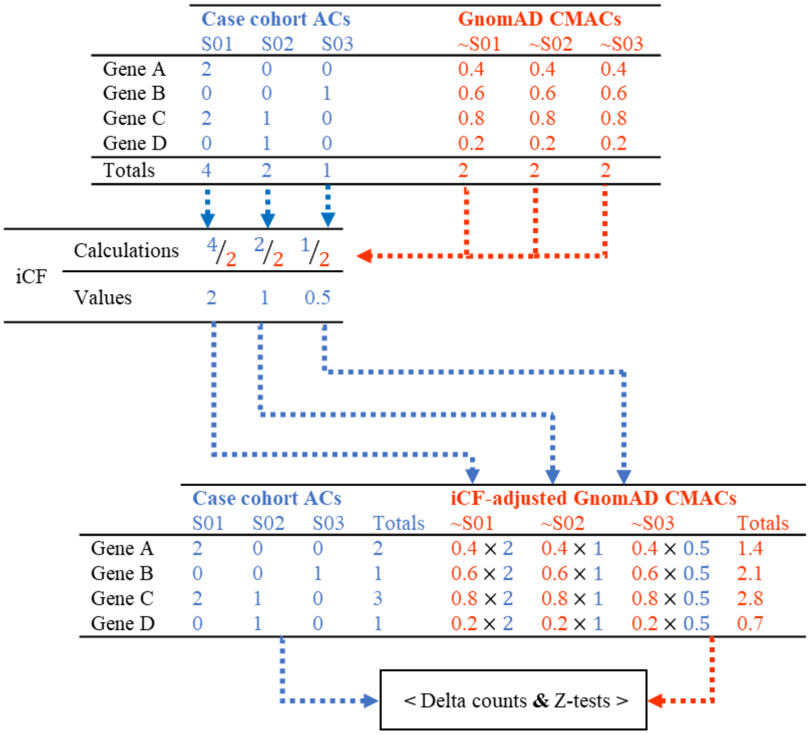
2.5.5 RV EXCALIBER

All samples in an exome sequencing study are typically selected from the same localized population and are sequenced with the same technology. This methodological consistency is lost when comparing a given cohort to GnomAD, where population and methodological factors confound analyses [18]. Such inter-cohort differences in sequencing methodology and population ancestry were addressed using Rare Variant Exome CALIBration using External Repositories (RV EXCALIBER), which employed an individual correction factor (iCF) and a gene correction factor (gCF) [18]. The iCF accounted for variation in population substructure and sequencing chemistry by comparing mutation load across individuals [18]:

$$iCF = \frac{case\ exome\text{–}wide\ mutation\ load}{control\ exome\text{–}wide\ mutation\ load}$$

An iCF unique to each individual was multiplied with GnomAD's CMAC gene values, resulting in iCF-adjusted expected counts. New delta counts and z-tests were done using the original observed T2D case counts and the iCF-adjusted expected counts. A simplified example of iCF adjustment is shown in Figure 2.6.

Like the iCF, the gCF also compared case and control mutation load, but instead acted on groups of genes based on their mutability. The gCF was constructed using a third, independent calibration cohort to capture gene-specific mutation biases [18]. For each study cohort, the samples without T2D were used for gCF calibration. This was ideal because within the same study, the non-T2D samples were sequenced on the same platform as the T2D samples and thus subject to the same confounding factors



	Case cohort ACs			GnomAD CMACs		
	S01	S02	S03	~S01	~S02	~S03
Gene A	2	0	0	0.4	0.4	0.4
Gene B	0	0	1	0.6	0.6	0.6
Gene C	2	1	0	0.8	0.8	0.8
Gene D	0	1	0	0.2	0.2	0.2
Totals	4	2	1	2	2	2

iCF	Calculations	$^4/_2$	$^2/_2$	$^1/_2$
	Values	2	1	0.5

	Case cohort ACs				iCF-adjusted GnomAD CMACs			
	S01	S02	S03	Totals	~S01	~S02	~S03	Totals
Gene A	2	0	0	2	0.4 × 2	0.4 × 1	0.4 × 0.5	1.4
Gene B	0	0	1	1	0.6 × 2	0.6 × 1	0.6 × 0.5	2.1
Gene C	2	1	0	3	0.8 × 2	0.8 × 1	0.8 × 0.5	2.8
Gene D	0	1	0	1	0.2 × 2	0.2 × 1	0.2 × 0.5	0.7

Figure 2.6: iCF adjustment example. iCF is calculated separately for each individual and multiplied by the GnomAD expected counts.

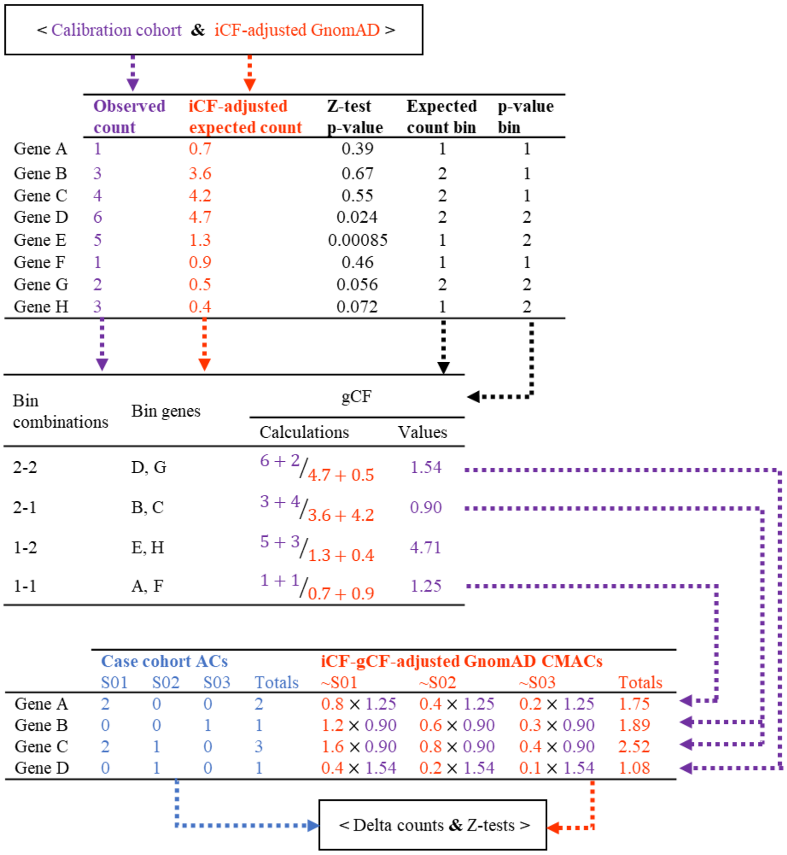
.

Following rare variant association methodology identical to that used on the case cohorts, the calibration cohorts underwent gCF adjustment. The iCF-adjusted genes in the calibration cohorts were then sorted and grouped: five quantiles of ascending expected counts and ten descending deciles of z-test p-values. The groups were then combined into 50 distinct bins and a gCF was calculated for each:

$$gCF = \frac{calibration\ bin\text{–}wide\ mutation\ load}{control\ bin\text{–}wide\ mutation\ load}$$

Genes in the case T2D cohorts were then matched to the bins and given the bin's gCF value. These gCF values were multiplied with the iCF-adjusted expected counts of

GnomAD, resulting in both iCF and gCF adjusted expected counts. A new, final set of delta counts and z-tests were conducted using the original observed T2D case counts and the iCF-gCF-adjusted expected counts. Figure 2.7 continues the simplified example with gCF adjustment using a calibration cohort.



< Calibration cohort & iCF-adjusted GnomAD >

	Observed count	iCF-adjusted expected count	Z-test p-value	Expected count bin	p-value bin
Gene A	1	0.7	0.39	1	1
Gene B	3	3.6	0.67	2	1
Gene C	4	4.2	0.55	2	1
Gene D	6	4.7	0.024	2	2
Gene E	5	1.3	0.00085	1	2
Gene F	1	0.9	0.46	1	1
Gene G	2	0.5	0.056	2	2
Gene H	3	0.4	0.072	1	2

Bin combinations	Bin genes	gCF Calculations	gCF Values
2-2	D, G	$^{6+2}/_{4.7+0.5}$	1.54
2-1	B, C	$^{3+4}/_{3.6+4.2}$	0.90
1-2	E, H	$^{5+3}/_{1.3+0.4}$	4.71
1-1	A, F	$^{1+1}/_{0.7+0.9}$	1.25

	Case cohort ACs				iCF-gCF-adjusted GnomAD CMACs			
	S01	S02	S03	Totals	~S01	~S02	~S03	Totals
Gene A	2	0	0	2	0.8 × 1.25	0.4 × 1.25	0.2 × 1.25	1.75
Gene B	0	0	1	1	1.2 × 0.90	0.6 × 0.90	0.3 × 0.90	1.89
Gene C	2	1	0	3	1.6 × 0.90	0.8 × 0.90	0.4 × 0.90	2.52
Gene D	0	1	0	1	0.4 × 1.54	0.2 × 1.54	0.1 × 1.54	1.08

< Delta counts & Z-tests >

Figure 2.7: gCF adjustment example. gCF is calculated for each gene and them multiplied by the GnomAD expected counts.

2.5.6 DBGap Meta-analysis

Like with the intracohort analyses, an inter-cohort version of the pooled meta-analysis of the three DBGap studies was also done. Each cohort underwent RV-

EXCALIBER correction separately, then gene delta counts from each cohort were added together. Genes with delta counts in only one of the three studies were excluded. Total sample sizes were adjusted accordingly for each remaining gene.

2.6 Summary

The methodologies covered in this chapter included reviews of exome sequencing protocols, selection criteria for T2D cases and controls, and two approaches for RV gene burden analysis of T2D risk. Following quality control and filtering for protein-altering variants, in the intracohort analyses RV burden gene was compared in T2D cases and controls of each study. With additional establishment of non-differential coverage and use of RV-EXCALIBER, the inter-cohort analyses compared RV gene burden of T2D cases to GnomAD as a summary level control.

2.7 References

1. Head, S. R., Komori, H. K., LaMere, S. A., Whisenant, T., Van Nieuwerburgh, F., Salomon, D. R., & Ordoukhanian, P. (2014). Library construction for next-generation sequencing: overviews and challenges. *BioTechniques*, *56*(2), 61–passim. https://doi.org/10.2144/000114133

2. Damiati, E., Borsani, G., & Giacopuzzi, E. (2016). Amplicon-based semiconductor sequencing of human exomes: performance evaluation and optimization strategies. *Human genetics*, *135*(5), 499–511. https://doi.org/10.1007/s00439-016-1656-8

3. Ivády, G., Madar, L., Dzsudzsák, E., Koczok, K., Kappelmayer, J., Krulisova, V., Macek, M., Jr, Horváth, A., & Balogh, I. (2018). Analytical parameters and validation of homopolymer detection in a pyrosequencing-based next generation sequencing system. *BMC genomics*, *19*(1), 158. https://doi.org/10.1186/s12864-018-4544-x

4. Kim, K., Seong, M. W., Chung, W. H., Park, S. S., Leem, S., Park, W., Kim, J., Lee, K., Park, R. W., & Kim, N. (2015). Effect of Next-Generation Exome Sequencing Depth for Discovery of Diagnostic Variants. *Genomics & informatics*, *13*(2), 31–39. https://doi.org/10.5808/GI.2015.13.2.31

5. Dou, J., Sun, B., Sim, X., Hughes, J. D., Reilly, D. F., Tai, E. S., Liu, J., & Wang, C. (2017). Estimation of kinship coefficient in structured and admixed populations using sparse sequencing data. *PLoS genetics*, *13*(9), e1007021. https://doi.org/10.1371/journal.pgen.1007021

6. Jong, S., Kim, J., Park, W., Jeon, H., & Kim, N. (2017). SEXCMD: Development and validation of sex marker sequences for whole-exome/genome and RNA sequencing. *PloS one*, *12*(9), e0184087. https://doi.org/10.1371/journal.pone.0184087

7. Privacy Notice for UK Biobank Participants. (2021). *The UK Biobank*. Retrieved November 3 2020, from: https://www.ukbiobank.ac.uk/withdrawal/

8. Lynch, M., Bost, D., Wilson, S., Maruki, T., & Harrison, S. (2014). Population-genetic inference from pooled-sequencing data. *Genome biology and evolution*, *6*(5), 1210–1218. https://doi.org/10.1093/gbe/evu085

9. Abramovs, N., Brass, A., & Tassabehji, M. (2020). Hardy-Weinberg Equilibrium in the Large Scale Genomic Sequencing Era. *Frontiers in Genetics*, *11*. https://doi.org/10.3389/fgene.2020.00210

10. Tan, A., Abecasis, G. R., & Kang, H. M. (2015). Unified representation of genetic variants. *Bioinformatics (Oxford, England)*, *31*(13), 2202–2204. https://doi.org/10.1093/bioinformatics/btv112

11. O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C. M., … Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic acids research*, *44*(D1), D733–D745. https://doi.org/10.1093/nar/gkv1189

12. Katsonis, P., Koire, A., Wilson, S. J., Hsu, T. K., Lua, R. C., Wilkins, A. D., & Lichtarge, O. (2014). Single nucleotide variations: biological impact and theoretical interpretation. *Protein science : a publication of the Protein Society*, *23*(12), 1650–1666. https://doi.org/10.1002/pro.2552

13. Pagel, K. A., Pejaver, V., Lin, G. N., Nam, H. J., Mort, M., Cooper, D. N., Sebat, J., Iakoucheva, L. M., Mooney, S. D., & Radivojac, P. (2017). When loss-of-function is loss of function: assessing mutational signatures and impact of loss-of-function genetic variants. *Bioinformatics (Oxford, England)*, *33*(14), i389–i398. https://doi.org/10.1093/bioinformatics/btx272

14. Kennedy, D. A., & Dwyer, G. (2018). Effects of multiple sources of genetic drift on pathogen variation within hosts. *PLoS biology*, *16*(3), e2004444. https://doi.org/10.1371/journal.pbio.2004444

15. Jagadeesh K.A.,, Wenger A.M., Berger M.J., Guturu H., Stenson P.D., Cooper D.N., Bernstein J.A., & Bejerano G. (2016). M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nature Genetics*, 48(12):1581-1586.

16. Fadista, J., Manning, A., Florez, J. *et al.* (2016). The (in)famous GWAS *P*-value threshold revisited and updated for low-frequency variants. *Eur J Hum Genet 24*, 1202–1205. https://doi.org/10.1038/ejhg.2015.269

17. Sanghvi, R. V., Buhay, C. J., Powell, B. C., Tsai, E. A., Dorschner, M. O., Hong, C. S., Lebo, M. S., Sasson, A., Hanna, D. S., McGee, S., Bowling, K. M., Cooper, G. M., Gray, D. E., Lonigro, R. J., Dunford, A., Brennan, C. A., Cibulskis, C., Walker, K., Carneiro, M. O., Sailsbery, J., … NHGRI Clinical Sequencing Exploratory Research (CSER) Consortium (2018). Characterizing reduced coverage regions through comparison of exome and genome sequencing data across 10 centers. *Genetics in medicine : official journal of the American College of Medical Genetics*, *20*(8), 855–866. https://doi.org/10.1038/gim.2017.192

18. Lali, R., Chong, M., Omidi, A., Mohammadi-Shemirani, P., Le, A., & Paré, G. (2020). Calibrated rare variant genetic risk scores for complex disease prediction using large exome sequence repositories. *BioRxiv*, 2020.02.03.931519. https://doi.org/10.1101/2020.02.03.931519

19. Type 2 Diabetes Genetic Exploration by Next-Generation Sequencing in Multi-Ethnic Samples (T2D-GENES) Project 1: KARE. (2019). *DBGap*. phs001096.v1.p1. Retrieved January 20, 2020 from: https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001096.v1.p1

20. Type 2 Diabetes Genetic Exploration by Next-Generation Sequencing in Multi-Ethnic Samples (T2D-GENES) Project 1: Metabolic Syndrome in Men Study (METSIM). (2019). *DBGap*. phs001100.v1.p1. Retrieved January 20, 2020 from: https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001100.v1.p1

21. T2D-GENES Consortium: San Antonio Mexican American Family Studies (SAMAFS). (2019). *DBGaP*. phs000847.v2.p1. Retrieved January 20, 2020 from: https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000847.v1.p

22. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ & Sham PC. (2007). PLINK: a toolset for whole-genome association and population-based linkage analysis. *American Journal of Human Genet*ics, 81.

23. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

24. Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

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

26. Garrison, E., Kronenberg, Z. N., Dawson, E. T., Pedersen, B. S., & Prins, P. (2021). Vcflib and tools for processing the VCF variant call format. *BioRxiv*, 2021.05.21.445151. https://doi.org/10.1101/2021.05.21.445151

27. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. (2002). The human genome browser at UCSC. *Genome Res.* 12(6):996-1006.

28. Haeussler, M., Zweig, A. S., Tyner, C., Speir, M. L., Rosenbloom, K. R., Raney, B. J., Lee, C. M., Lee, B. T., Hinrichs, A. S., Gonzalez, J. N., Gibson, D., Diekhans, M., Clawson, H., Casper, J., Barber, G. P., Haussler, D., Kuhn, R. M., & Kent, W. J. (2019). The UCSC Genome Browser database: 2019 update. *Nucleic acids research*, *47*(D1), D853–D858. https://doi.org/10.1093/nar/gky1095

29. Wang K, Li M, Hakonarson H. (2010). ANNOVAR: Functional annotation of genetic variants from next-generation sequencing data. *Nucleic Acids Research,* 38:e164.

30. Tsalamandris, S., Antonopoulos, A. S., Oikonomou, E., Papamikroulis, G. A., Vogiatzi, G., Papaioannou, S., Deftereos, S., & Tousoulis, D. (2019). The Role of Inflammation in Diabetes: Current Concepts and Future Perspectives. *European cardiology*, *14*(1), 50–59. https://doi.org/10.15420/ecr.2018.33.1

31. Danecek, P., Auton, A., Abecasis, G., Albers, CA., Banks, E., DePristo, MA., Handsaker, R., Lunter, G., Marth, G., Sherry, ST., McVean, G., Durbin, R. & 1000 Genomes Project Analysis Group. (2011). *The Variant Call Format and VCFtools, Bioinformatics*.

32. Choi, S. H., Jurgens, S. J., Weng, L. C., Pirruccello, J. P., Roselli, C., Chaffin, M., Lee, C. J., Hall, A. W., Khera, A. V., Lunetta, K. L., Lubitz, S. A., & Ellinor, P. T. (2020). Monogenic and Polygenic Contributions to Atrial Fibrillation Risk: Results From a National Biobank. Circulation research, 126(2), 200–209. https://doi.org/10.1161/CIRCRESAHA.119.315686

33. Khera, A. V., Chaffin, M., Aragam, K. G., Haas, M. E., Roselli, C., Choi, S. H., Natarajan, P., Lander, E. S., Lubitz, S. A., Ellinor, P. T., & Kathiresan, S. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature genetics*, *50*(9), 1219–1224. https://doi.org/10.1038/s41588-018-0183-z

34. Fadista, J., Manning, A., Florez, J. *et al.* (2016). The (in)famous GWAS *P*-value threshold revisited and updated for low-frequency variants. *Eur J Hum Genet*. 24, 1202–1205. https://doi.org/10.1038/ejhg.2015.269

35. Slatkin M. (2008). Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. *Nature reviews. Genetics*, *9*(6), 477–485. https://doi.org/10.1038/nrg2361

36. Turkmen A, Lin S. (2017). Are rare variants really independent? *Genet Epidemiol*. 41(4):363-371. doi: 10.1002/gepi.22039.

37. Hunt, K. J., Lehman, D. M., Arya, R., Fowler, S., Leach, R. J., Göring, H. H. H., Almasy, L., Blangero, J., Dyer, T. D., Duggirala, R., & Stern, M. P. (2005). Genome-Wide Linkage Analyses of Type 2 Diabetes in Mexican Americans. Diabetes, 54(9), 2655 LP – 2662. https://doi.org/10.2337/diabetes.54.9.2655

38. Accounting for sex in the genome. (2017). *Nat Med*, 23, 1243. https://doi.org/10.1038/nm.4445

39. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., Cortes, A., Welsh, S., McVean, G., Leslie, S., Donnelly, P., & Marchini, J. (2017). Genome-wide genetic data on ~500,000 UK Biobank participants. BioRxiv, 166298. https://doi.org/10.1101/166298

40. Otto, P. A., & Horimoto, A. R. (2012). Penetrance rate estimation in autosomal dominant conditions. Genetics and molecular biology, 35(3), 583–588. https://doi.org/10.1590/S1415-47572012005000051

41. Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

CHAPTER 3 – Results

3.1 Overview

The results presented in chapter are divided into four main sections: highlights of top intracohort rare variant (RV) gene associations, illustrations of genomic inflation, and assessments of the inter-cohort correction factors. All other result data are provided with the supplementary materials.

3.2 RV Gene Burden Associations

Due to the number of varied analyses, RV association results are presented by using the most well powered and calibrated comparison, Type 2 Diabetes (T2D) cases versus controls in the United Kingdom Biobank (UKB), as a discovery cohort. Table 3.1 showcases the top genes from this analysis and includes the number of carriers, the estimated percent penetrance of each gene, the odds ratios for the risk of T2D per unit increase of RV gene burden, and the gene association P-values.

Table 3.1: Top genes in intracohort analysis in the UKB

Gene	**Number of RV carriers**	**Percent Penetrance**	**Odds Ratio (95% CIs)**	**P-value**
GCK	509	11.98	2.44 (1.86 - 3.20)	8.91×10^{-11}
PAM	4784	6.84	1.32 (1.18 - 1.48)	1.39×10^{-6}
FGF16	78	14.10	4.03 (2.11 - 7.70)	2.55×10^{-5}
HNF4A	1123	8.28	1.55 (1.25 - 1.93)	6.82×10^{-5}
NEPRO	852	8.45	1.63 (1.28 - 2.08)	7.06×10^{-5}
TM4SF20	635	8.50	1.74 (1.32 - 2.29)	8.33×10^{-5}
BTF3	100	15.00	3.19 (1.78 - 5.69)	9.06×10^{-5}
ACE	8936	6.21	1.18 (1.08 - 1.29)	1.33×10^{-4}
KAT14	1008	7.84	1.50 (1.21 - 1.86)	1.77×10^{-4}
TBKBP1	2198	7.05	1.36 (1.16 - 1.60)	1.84×10^{-4}
IER2	179	11.73	2.40 (1.50 - 3.83)	2.39×10^{-4}
RAI2	453	7.28	1.90 (1.34 - 2.70)	2.95×10^{-4}

RAD52	217	10.60	2.22 (1.44 - 3.42)	3.23×10^{-4}
HTR1E	134	12.69	2.59 (1.54 - 4.36)	3.26×10^{-4}
ADGRD1	4925	6.40	1.23 (1.10 - 1.38)	3.58×10^{-4}
ZNF708	2702	6.55	1.31 (1.13 - 1.53)	3.87×10^{-4}
ZNF620	89	14.61	2.97 (1.63 - 5.43)	3.94×10^{-4}
NSUN4	476	9.66	1.76 (1.28 - 2.41)	4.26×10^{-4}
C1orf195	52	15.38	3.54 (1.74 - 7.22)	5.00×10^{-4}
RNMT	148	11.49	2.39 (1.46 - 3.90)	5.24×10^{-4}
ZNF184	108	12.04	2.73 (1.54 - 4.83)	5.86×10^{-4}
PTPN3	2075	7.18	1.35 (1.14 - 1.59)	5.94×10^{-4}
PPIAL4C	4581	4.39	0.81 (0.72 - 0.92)	5.94×10^{-4}
MCCC1	2153	6.69	1.32 (1.13 - 1.56)	6.63×10^{-4}
PCNX1	3590	6.52	1.26 (1.10 - 1.44)	6.86×10^{-4}
SYPL2	2415	6.83	1.31 (1.12 - 1.54)	7.01×10^{-4}
HEATR5B	993	7.65	1.39 (1.15 - 1.68)	7.28×10^{-4}
ABI1	1908	3.98	0.67 (0.53 - 0.84)	7.38×10^{-4}
CHML	943	8.06	1.50 (1.19 - 1.91)	7.60×10^{-4}
SLC45A2	2059	6.80	1.34 (1.13 - 1.59)	8.07×10^{-4}
GNPTAB	4757	6.24	1.21 (1.08 - 1.35)	8.51×10^{-4}
KCNK17	1748	7.55	1.37 (1.14 - 1.64)	8.79×10^{-4}
GSAP	414	8.94	1.77 (1.26 - 2.47)	8.91×10^{-4}
AMMECR1L	14	28.57	7.84 (2.32 - 26.52)	9.18×10^{-4}
ABAT	1302	3.23	0.59 (0.43 - 0.81)	9.32×10^{-4}

Genes with $p < 0.001$ in intracohort analysis of RV burden in the UKB. The dashed line marks the exome wide significance threshold for multiple hypothesis testing ($p<2.65\times10^{-6}$).

Of the top genes, only two had replication of significant RV association with T2D in either intracohort or inter-cohort analyses of the DBGap cohorts: *ACE* in Korean Association Resource (KARE) project cases versus the Genome Aggregation Database (GnomAD), (OR = 1.46, P = 0.021) and *TBKBP1* in the Metabolic Syndrome in Men Study (METSIM), cases versus controls (OR = 2.38, P = 0.030). There was no replication of top genes in the San Antonio Mexican American Family Studies (SAMAFS), or in the meta-analysis of KARE, METSIM, and SAMAFS (KMS). Several genes had exome-wide significant associations with T2D in the analysis of UKB cases versus GnomAD (Table S6).

Two strongly associated genes in the RV burden analysis, *GCK* and *HNF4A*, are known to cause 30-50% and 5-10% of cases in Mature Onset Diabetes of the Young (MODY), respectfully [1]. Despite being responsible for a majority 30-65% of MODY cases *HNF1A* was not significant in the intracohort UKB analysis [1]. *ABCC8, BLK*, and *CEL*, each a minor MODY representing under 1% of cases, showed some significance (P < 0.05) in both the intracohort and intercohort analyses of the UKB, METSIM, and KMS (Table 3.2).

Table 3.2: Associations of MODY genes with T2D

Cohort	**UKB**		**KARE**		**METSIM**		**SAMAFS**		**KMS**	
Intracohort Analyses: T2D cases vs. controls										
Gene	**OR**	**P**	**OR**	**P**	**OR**	**P**	**OR**	**P**	**OR**	**P**
HNF1A	1.06	0.49	1.32	0.64	1.53	0.34	0.73	0.63	1.32	0.31
GCK	2.44	8.90×10^{-11}	0.56	0.73	0.37	0.55			1.49	0.60
HNF4A	1.55	6.82×10^{-5}	3.56	0.07	0.60	0.54			1.14	0.76
HNF1B	1.03	0.76	0.61	0.24	0.86	0.77	0.30	0.089	0.90	0.63
PDX1	1.07	0.27	0.67	0.66	2.20	0.24			2.21	0.051
ABCC8	1.19	0.0033	0.98	0.95	0.41	0.011	1.40	0.49	0.78	0.12
APPL1	0.83	0.25	2.64	0.50					7.22	0.10
BLK	1.13	0.034	0.23	0.24	5.57	0.00021	2.44	0.31	2.45	0.015
CEL	0.94	0.17	0.96	0.74	1.60	0.017	0.81	0.61	1.13	0.10
INS	0.91	0.81			0.79	0.80				
KCNJ11	1.31	0.06	3.04	0.53	1.21	0.90	0.64	0.72	0.74	0.69
KLF11	1.15	0.13	4.87	0.40						
NEUROD1	0.72	0.16	0.76	0.75					1.22	0.74
PAX4	1.02	0.72	0.34	0.14	1.51	0.47	0.79	0.85	1.04	0.90
Inter-cohort Analyses: T2D cases vs. GnomAD										
Gene	**OR**	**P**	**OR**	**P**	**OR**	**P**	**OR**	**P**	**OR**	**P**
HNF1A	1.10	0.17	1.00	0.50			1.13	0.41	1.05	0.44
GCK	1.13	0.24								
HNF4A	1.62	0.00034								
HNF1B	0.92	0.75	0.39	1.00						
PDX1	1.04	0.29								
ABCC8	1.15	0.029	1.20	0.20	0.51	0.98	2.30	0.053	1.11	0.26
APPL1	0.97	0.59								
BLK	1.12	0.029	0.40	0.98	2.77	0.0060	1.48	0.23	1.75	0.022
CEL	1.17	0.010	1.24	0.24	5.15	0.0022			2.10	0.0042
INS										
KCNJ11	1.33	0.049			0.27	0.99				
KLF11	1.28	0.054								
NEUROD1	0.59	0.99								
PAX4	1.07	0.24	0.98	0.52						

Odds ratios (ORs) and P-values (P) of all 14 major and minor MODY genes, separated by intracohort and intercohort analyses.

30 of the 35 genes in Table 3.1 had previously been identified in genome wide association studies (GWAS) to have significant associations with other phenotypes. These associations were largely based on common variants located anywhere within the boundaries of each gene, including non-coding regions like promoters or introns. The phenotype that had the strongest association with common variants in each gene are shown in Table 3. Some genes contained common variants that were associated with numerous different phenotypes. The total number of phenotypes that have reached genome wide significance ($P < 5\times10^{-8}$) with common variants in the genes are also shown. Its notable that the strongest associated phenotypes for several of the genes was T2D or a related phenotype, suggesting a shared functional impact of RVs and common variants in those genes.

Table 3.3: Genes with previous associations found in GWAS

Gene	Strongest phenotype association	Genome-wide significant phenotypes
GCK	Random glucose	16
ZNF184	Waist-hip ratio adj BMI	14
HNF4A	Type 2 Diabetes	11
SYPL2	LDL cholesterol	10
PAM	Type 2 diabetes adj BMI	9
ACE	Height	8
TBKBP1	LDL cholesterol	8
HEATR5B	PR interval	7
IER2	Height	6
PTPN3	Diastolic blood pressure	6
KCNK17	Type 2 diabetes	6
TM4SF20	Late diabetic kidney disease adj HbA1c-BMI	5
RAI2	Urinary albumin-to-creatinine ratio	5

RAD52	BMI	4
PCNX1	PR interval	4
GSAP	BMI	4
FGF16	Urinary albumin	3
NEPRO	Height	2
KAT14	BMI	2
ADGRD1	Height	2
BTF3	Height	1
HTR1E	BMI	1
ZNF708	Type 2 diabetes	1
ZNF620	PR interval	1
NSUN4	Waist-hip ratio	1
RNMT	Height	1
ABI1	BMI	1
SLC45A2	Any cancer	1
GNPTAB	Height	1
ABAT	Height	1

All fields were obtained from respective gene pages on the Type 2 Diabetes Knowledge Portal [3]. BMI is body mass index. LDL is low density lipoprotein.

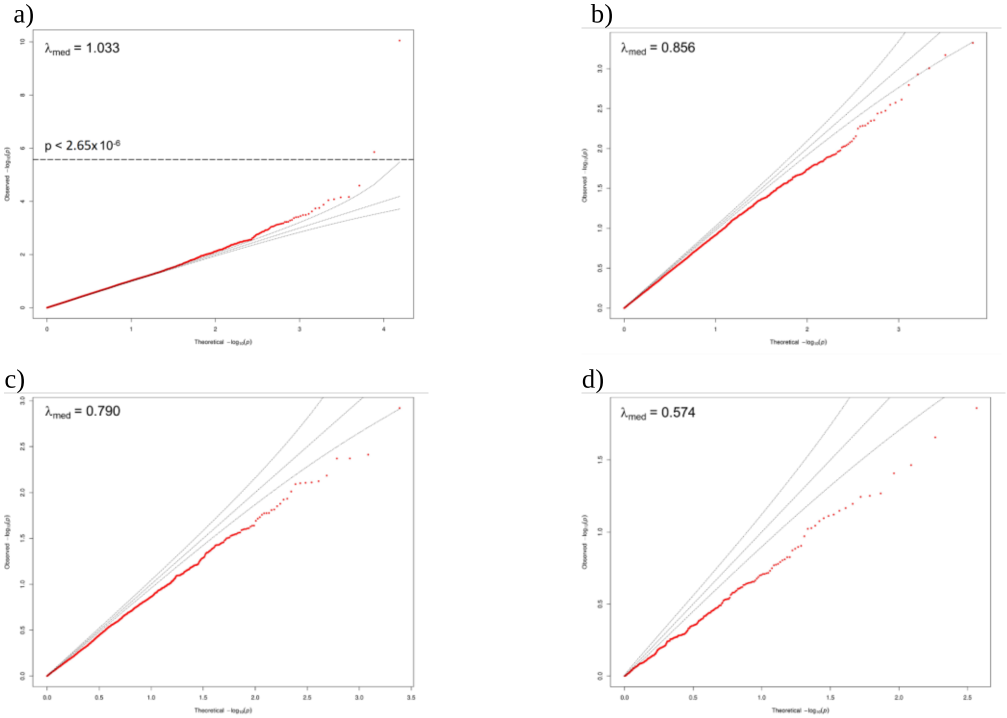
3.3 Distributions of Gene Significance

Gene p-value distributions of all intracohort (Figure 3.1) and inter-cohort (Figure 3.2) analyses are displayed on quantile-quantile (QQ) plots with calculated genomic inflation factors (λ_{med}). Gene association P-values (P) are first converted to P_{new}:

$$P_{new} = -\log_{10} P$$

Which effectively inverses the P-value, making small, statistically significant P-values into large values that are easier to visualize on the QQ plot. Then, these converted P-values (observed) are ordered by increasing significance and are plotted against an equal number of ordered P-values (expected) that follow a uniform distribution. Most genes in the exome are not expected to be functionally related to T2D, so the distribution of observed association P-values should closely match the uniform distribution of the expected P-values, resulting in a QQ plot with a straight diagonal line. Genes that are truly related to T2D should have P-values of great enough significance that they appear above the diagonal line and its 95% confidence interval wings on the QQ plot. However, if most genes are above the diagonal line on the QQ plot, they may be false positives and indicate systematic genomic inflation. In this context, genomic inflation refers to overestimation of statistical significance due to RV gene burden. Conversely, if most genes are below the diagonal, then they may be false negatives and indicate systematic genomic deflation. QQ plots do not reveal the causes of the inflation or deflation but are good visual indicators for them.

Genomic inflation can also be calculated as λ_{med}, which simply the median of the observed P-values divided by the median of the expected P-values. Genomic inflation is indicated when $\lambda_{med} > 1$ and deflation when $\lambda_{med} < 1$.



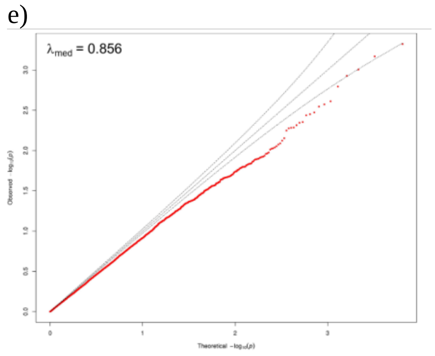
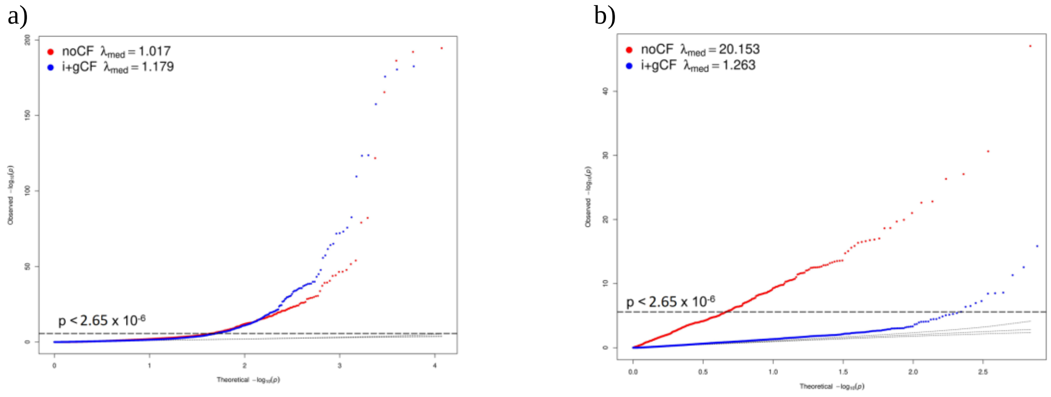


Figure 3.1: Gene-based QQ plots for intracohort analyses. Corresponding study cohorts include a) the UKB, b) KARE, c) METSIM, d) SAMAFS, and d) KMS meta-analysis. All plots include genes with case MAC >= 10 and control MAC >= 10. The dashed line in a) marks the exome wide significance threshold for multiple hypothesis testing (p<2.65×10-6).

GCK and *PAM* show up in their expected spots in the upper righthand corner of Figure 3.1a as their exome-wide significance exceeded the expected distribution. Genomic inflation in genes of the intracohort UKB analysis appears to be minor both visually and by a λ_{med} close to 1. However, the other intracohort analyses in depicted in Figures 3.1$_{b-e}$ show moderate genomic deflation and is most severe in the smallest cohort SAMAFS.



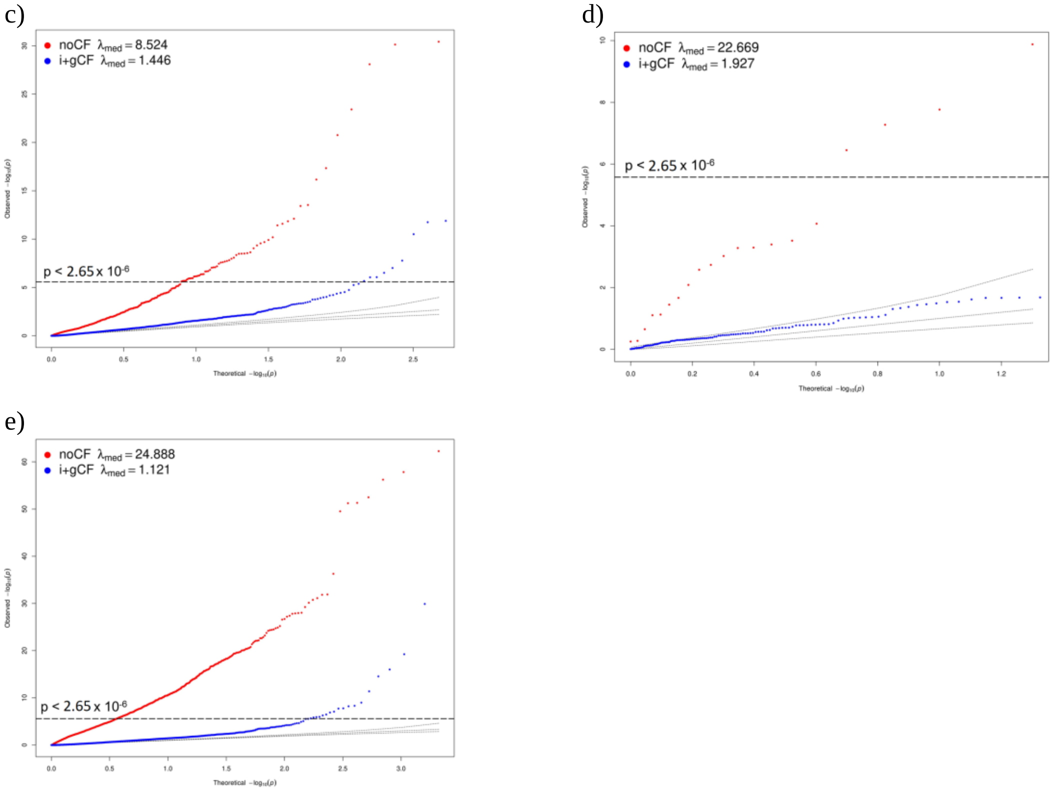


Figure 3.2: Gene-based QQ plots for inter-cohort analyses. Dots denoted as before (red) and after (blue) individual & gene level correction with RV-EXCALIBER. Corresponding study cohorts: a) the UKB, b) KARE, c) METSIM, d) SAMAFS, and d) KMS meta-analysis. All plots include genes with case MAC >= 10 and GnomAD CMAC => 10. The dashed line marks the exome wide significance threshold for multiple hypothesis testing (p<2.65×10-6).

Genomic inflation was much more pronounced in the intercohort analyses. However, it was dramatically decreased following correction with RV-EXCALIBER in all cohorts except the UKB. Post correction, exome-wide significant genes appeared to be present in every cohort except SAMAFS.

3.4 Correction Factor Performance

Differences between individual correction factor (iCF) and gene correction factor (gCF) calibration in RV-EXCALIBER were demonstrated with folded cumulative distribution function plots for delta count plots, also called mountain plots (Figure 3.3). Gene-level bias in iCF correction was represented by inflation of observed counts relative to GnomAD, followed by deflation, forming a "mountain". Correction of this bias with the gCF "flattens" the mountain.

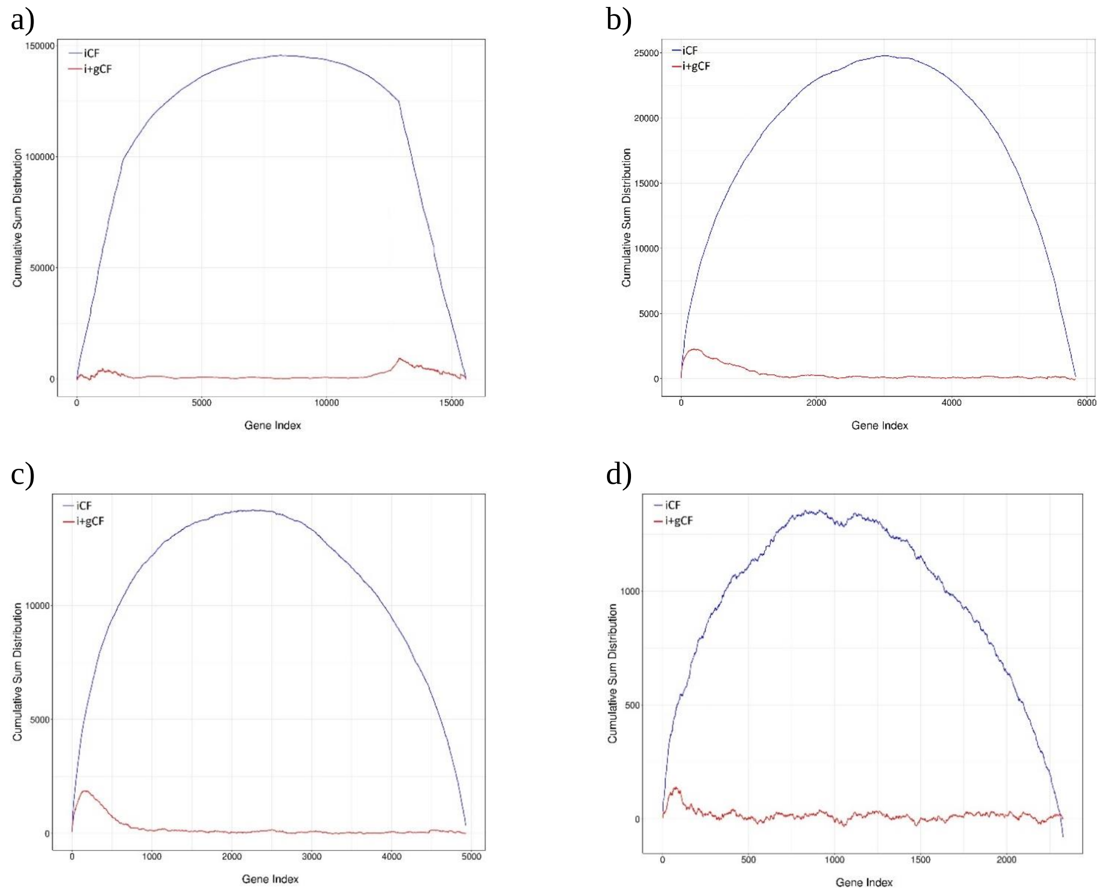


Figure 3.3: Mountain plots of cumulative gene delta count distribution. Gene index was ordered by z-test p-value. Corresponding study cohorts: a) the UKB, b) KARE, c) METSIM, and d) SAMAFS. Correction factor calculations in all plots included genes with observed case MAC >=1 and expected GnomAD CMAC >=1.

Across all cohorts, gene-level bias remaining after iCF correction appeared to be successfully corrected following gCF correction.

3.5 Supplementary Materials

Please see the attached spreadsheet containing tables of all RV gene burden associations across both intracohort and inter-cohort analyses (S1-S10).

3.6 Summary

In the intracohort analysis of T2D risk in the UKB, RV burden in two genes reached exome wide significance, *GCK* (OR = 2.44, P = 8.91×10^{-11}) and *PAM* (OR = 1.32, P = 1.39×10^{-6}), though neither was replicated in the other cohorts. In addition to *GCK*, several other genes implicated in the monogenic diabetes subtype MODY showed suggestive significance in multiple cohorts: *HNF4A*, *ABCC8*, *BLK*, and *CEL*. Many top genes also contained common variants with previously established associations with T2D or related phenotypes. Examination of P-values of all genes showed a lack of genomic inflation in the intracohort analyses, though there was deflation in KARE, METSIM, and SAMAFS. Genomic inflation was present in all inter-cohort analyses but was reduced after correction with RV-EXCALIBER.

3.7 References

1. Naylor R, Knight Johnson A, del Gaudio D. (2018), Maturity-Onset Diabetes of the Young Overview. (Adam MP, Ardinger HH, Pagon RA, et al., editors). *GeneReviews* [Internet]. https://www.ncbi.nlm.nih.gov/books/NBK500456/

2. de Leeuw C, Mooij J, Heskes T,& Posthuma D. (2015). MAGMA: Generalized gene-set analysis of GWAS data. PLoS Comput Biol, *11*(4), e1004219. doi:10.1371/journal.pcbi.1004219

3. Type 2 Diabetes Knowledge Portal. (2021). Providing data and tools to promote understanding and treatment of type 2 diabetes and its complications. Retrieved January 25 2020 from: https://t2d.hugeamp.org/

4. Yang, J., Weedon, M. N., Purcell, S., Lettre, G., Estrada, K., Willer, C. J., Smith, A. V., Ingelsson, E., O'Connell, J. R., Mangino, M., Mägi, R., Madden, P. A., Heath, A. C., Nyholt, D. R., Martin, N. G., Montgomery, G. W., Frayling, T. M., Hirschhorn, J. N., McCarthy, M. I., Goddard, M. E., … GIANT Consortium (2011). Genomic inflation factors under polygenic inheritance. *European journal of human genetics : EJHG*, *19*(7), 807–812. https://doi.org/10.1038/ejhg.2011.39

CHAPTER 4 – Discussion

4.1 Overview

The discussion in this chapter first covered the results of RV gene burden associations with Type 2 Diabetes (T2D) from intracohort and intercohort analyses in the UK Biobank (UKB), the Korea Association Research Project (KARE), the Metabolic Syndrome Study in Men (METSIM), and the San Antonio Mexican American Family Studies (SAMAFS), and a meta-analysis of KARE, METSIM, SAMAFS (KMS). Discussion focused on the clinical significance of the top gene *GCK*, as well as exploring research implications of *PAM* and other suggestive genes, then follow-up analyses were suggested. The discussion moves onto examining limitations in the methodology and proposing improvements.

4.2 RV Gene Burden Results

4.2.1 Clinical Significance of GCK

The gene of greatest significance with T2D risk was *GCK*, a major casual gene of the rare monogenic subtype of diabetes called Mature Onset Diabetes of the Young (MODY). Four other MODY genes, *HNF4A*, *ABCC8*, *BLK*, and *CEL*, also had suggestive significance in the UKB and in several intercohort analyses, suggesting the associations were driven by enrichment in T2D cases rather than just depletion in controls. However, it is notable that the gene *HNF1A*, which causes the most cases of MODY, did not have significant association in RV burden with T2D in any cohort or analysis. Distinct from Type 1 Diabetes (T1D) and T2D, MODY is characterized by an age of diagnosis (AOD) before 25 years, lack of autoantibodies, and an autosomal dominant mode of inheritance [8]. MODY accounts for up to 5% of all diabetes cases, with a population prevalence of about 1 in 10,000, though many cases go undiagnosed [8,69]. MODY has several subtypes, each characterized by mutations in different genes with specific effects on pancreatic β-cells [8]. The three most common MODY subtypes, describing 99% of cases, are caused by mutations in the hepatocyte nuclear factor-1-alpha (HNF1A) gene, followed by the glucokinase (GCK) and hepatocyte nuclear factor-4-alpha (HNF4A) genes [8]. HNF1A and HNF4A are involved in transcriptional regulation of the insulin gene, whereas *GCK* encodes a glucose sensor that can dictate the glycemic threshold for insulin secretion [8]. Eleven additional gene products have been associated with MODY, including ATP-binding cassette, subfamily C, member 8 (ABCC8), a nonreceptor tyrosine-kinase of the src family of proto-oncogen (BLK), and carboxyl ester lipase (CEL) [9,10,68]. *ABCC8* encodes a receptor for insulin-secretion-inducing sulfonylureas, BLK upregulates insulin transcription factors, and CEL digests dietary fats and fat-soluble vitamins in the small intestine [9,10,68]. About 90% of diabetes cases in the UKB had an AOD greater than 37 years, so MODY would have been an unlikely diagnosis in the cohort [11]. Still the enrichment of MODY gene mutations in UKB T2D cases suggest the two forms of diabetes may not be as genetically distinct as previously thought. Indeed, recent efforts have been made to redefine T2D into distinct subtypes by cluster analysis of clinical biomarkers and risk factors [12]. Exploring the genetic context of T2D heterogeneity could underscore this line of investigation.

4.2.2 Research Implications of PAM and other genes

The second strongest exome-wide significant association to T2D risk was in the gene that encodes the protein peptidylgycine α-amidating monooxygenase (PAM). PAM is not clinically recognized in T2D pathology, however recent evidence has suggested otherwise [13]. PAM is involved in the synthesis of neuropeptides, with a strong functional dependence on copper availability [14]. Therefore, PAM has been historically linked to the copper metabolism disorder Menke's disease [15]. While insulin is not a substrate of PAM, its expression in pancreatic islets suggests it has some functionality in β-cells [16]. In an experiment using human cell lines and islets taken from cadavers, Thomsen, *et al*., demonstrated that T2D risk factors caused PAM to lose functionality and lead to β-cell dysfunction [16]. They proposed that PAM mediates β-cell insulin secretion via granule exocytosis [16]. Missense mutations in *PAM* have previously been associated with T2D in GWAS, and recently, in the RV burden analysis by Flannick, *et al*. (OR = 1.31, P = 4.28 $\times 10^{-9}$) [13,16]. While KARE, METSIM, and SAMAFS were included in Flannick, *et al*.'s study, it is notable that *PAM* enrichment was not significant in our analysis of those cohorts. Carriers of protein altering mutations in *PAM* were present in all three cohorts: 13 in KARE, 3 in METSIM and 7 in SAMAFS (Tables S2, S3, and S4). The pooled cohort of the Database of Genotypes and Phenotypes (DBGap) and the European Genome-phenome Archive (EGA) studies utilized by Flannick, *et al*. was twenty times larger than our KMS meta-analysis [13]. This provided them sufficient power to detect significant *PAM* enrichment, albeit in an ethnically heterogeneous population [7,17].

Many genes with suggestive associations between RV burden and T2D contained common variants that had previously been identified to be significantly associated with several phenotypes (Table 3.3). Some of these previously associated phenotypes are T2D, lending credence to the observed association of T2D with RV burden in those genes. However, some of the phenotypes were not directly related to T2D. For example, common variants within *SYPL2* are strongly associated with low density lipoprotein (LDL) cholesterol, a major risk factor for coronary artery disease (CAD) [4,5]. In turn, those diagnosed with T2D have a greatly increased risk of developing CAD [1]. The shared presence of RVs associated with T2D and common variants associated with LDL cholesterol in *SYPL2* suggest the gene is involved in the pathology of both T2D and CAD. However, such speculation is tenuous because *SYPL2*s association with T2D did not reach exome wide significance and the functional impact of common and rare variants may differ despite proximity in a given gene region [7]. Still, this could justify a project testing risk of CAD or LDL cholesterol concentration between *SYPL2* carriers and non-carriers with T2D. The relationships between T2D and the other phenotypes identified in Table 3.3 may also be worth exploring.

4.2.3 Follow-up Analyses

In addition to specific follow-up analyses for mutations in MODY genes and *PAM*, there are two general avenues that should be considered.

First, its important to recognize that while common and rare variation act independently on T2D risk, they still both contribute to the measured outcome [7]. New polygenic risk scores derived from common variation in the UKB can be included as covariates to investigate their effect on the RV gene burden associations [28]. It is also possible to construct rare variant genetic risk scores (RVGRS) using the RV gene burden test statistics and test how well they predict T2D. RVGRS would be constructed in a sample by first weighting each gene's RV burden by its estimated effect size on T2D risk, then taking the sum of all genes that meet an optimized P-value threshold [33]. The ability of RVGRS to predict T2D can be compared between cohorts and contrasted with risk scores based on common variants or phenotypes.

Second, any findings can be further validated with replication in underrepresented non-European populations [19]. The current release of the UKB exomes includes about 12,000 non-European, unrelated individuals that could be utilized for this purpose. Further stratification of these 12,000 individuals into distinct, ethnically homogenous groups may jeopardize power to detect RV associations [17]. Aside from waiting for the full release of the 500,000 UKB exomes, sufficient sample size of these groups could be achieved by pooling each with ethnically corresponding DBGap cohorts.

Several options can be pursued to explore the relationship of MODY and T2D. Sensitivity tests can determine if explicit loss-of-function (frameshift, stop-gain, stop-loss, and start-loss) RVs in MODY genes were driving the observed associations. Adjustment for T2D medication use, such as metformin, can be easily checked as well. RV burden in a combination of MODY genes could be tested to see if they predict T2D risk more effectively than individual major or minor genes [7]. Inclusion of multiple genes would mean a larger genomic area in the comparison; increasing the power to detect RVs. This approach would be limited however by potentially expanded bidirectional effects between pathogenic and protective genes, an issue inherent to the RV gene burden model [17]. This issue, and a viable solution, are further discussed in Section 4.3. Another option would be to examine RV burden in MODY genes relative to biomarkers used to differentiate MODY from T2D or T1D, such as age of diagnosis (AOD), adiposity, triglyceride levels, or glutamic acid decarboxylase antibody levels [29,30]. In people with T2D, the concentration of these biomarkers could be tested between carriers and non-carriers of MODY RVs. The absence of a significant association of RV gene burden in *HNF1A* with T2D could be explored by redoing the analyses using T1D, as well as other diabetes subtypes, as the primary phenotype.

RV burden in *PAM* can be compared to measures of insulin secretion in carriers to test the effect on β-cell function [16]. Endogenous insulin secretion is measured by proxy using C-peptide, which is produced along with insulin in equal amounts from the cleavage of proinsulin [20]. Modern monoclonal based C-peptide assays are more affordable and reliable than traditional radioimmunoassays, however issues in test standardization have limited their adoption as a routine clinical measurement [20,21]. All three DBGap cohorts include C-peptide measurements, while the UKB does not at present. Though analysis may be possible in the future in 50,000 UKB samples if C-peptide is included in its upcoming panel of 1,500 circulating plasma proteins or its release of nuclear magnetic resonance assayed metabolomics [30,32].

4.3 Methodology Review

4.3.1 Limitations

The limitations of this thesis's methodology fall into four categories: statistical power in relation to cohort size and composition, selection criteria for samples and variants, assumptions of RV burden analysis, and requirements for implementing RV-EXCALIBER.

Replication of RV gene burden associations with T2D risk largely failed to replicate between the UKB and the three DBGap cohorts. While this trend of non-replication may have been partially explained by the ethnic diversity across the cohorts, it was most likely because KARE, METSIM, and SAMAFS were underpowered for RV burden association [17,19]. Using simulations of loci with known disease risk, Zhang, *et al*. estimated that at least 14,000 total cases and controls are needed to achieve a 90% rate of identifying true positives in burden analysis of variants with MAFs under 0.01 [17]. In contrast, the simulated true positive rate of cohorts the size of the DBGap studies, either individually or combined, would fall well short of 50% [17]. This is further illustrated by the actual number of nominally powered genes in each cohort. Out of approximately 20,000 unique protein coding genes, at least one RV was detected in both T2D cases and controls for: 18,815 in the UKB, 11,927 in KARE, 9,606 in METSIM, and 7,960 in SAMAFS (Supplementary Tables 1-4). While of sufficient total sample size for RV burden analysis, the UKB had its own unique issue: an unbalanced case to control ratio. The UKB has 8,784 T2D cases and 153,431 controls, a ratio of almost 1 to 20. Under these conditions, variant score statistics may deviate from the normal distribution and have higher false positive rates [39]. Though in Zhang, *et al*.'s simulations, they found that increasing the number of controls from 10,000 to 30,000, while keeping a constant ratio to cases, lowered the false positive rate [17]. With over 150,000 controls, the effects of an unbalanced case to control ratio in the UKB may be similarly subdued [40]. However, several techniques can reportedly control for these effects, including saddle point approximation (SPA) which was used by Jurgens, *et al*. in their recent RV burden study [41].

Jurgens, *et al*.'s study examined RV burden in UKB exomes with multiple phenotypes, including T2D. The differences between the methodology of this thesis and their sample and variant selection invite discussion. Their sample selection was more inclusive because they retained both related and non-European individuals for a total of 199,832 samples, which included 13,462 T2D cases [41]. While sample size was increased, their expanded sample selection criteria would have entailed some extra considerations. First, variant MAFs in related individuals are less dependent on their effect on disease risk and more so tied to familial inheritance [42]. RV burden analysis of a population that includes both related and unrelated individuals would result in inflated allele counts from the former. Though, as done by Jurgens, *et al*., a weighted adjustment could be used to account for relatedness [41]. The second consideration, inclusion of non-Europeans, has a consequence that is harder to quantify: increased ethnic heterogeneity. While RV associations seem to conform to specific ethnicities to a lesser degree compared to common variant associations, analysis including population stratification could still be insightful [7,18]. For example, people not of British ancestry make up about 11% of the entire UKB, but account for approximately 15% of all T2D cases (Table 1.3) [43]. RV burden may follow a likewise ununiform distribution.

With many thousands fewer samples, all unrelated and European, it is not unexpected that our T2D gene association results did not perfectly align with Jurgens, *et al.*'s. However, while we did overlap in the exome wide significance of *GCK*, their signal (OR = 13.9 [8.2-23.4], P = 1.46 $\times 10^{-19}$) was much stronger than ours [41]. They also found two novel associations in *GIGYF1* and *CCAR2*, which where non-significant in our analysis [41]. These discrepancies may be explained by differences in filtering criteria for protein altering variants and sample selection. Jurgens, *et al.* used a maximum MAF cut-off at 0.001 (0.1%) to filter for protein altering variants, while we were less stringent with 0.01 (1%) [41]. Also, where we only used MCAP at its default threshold, they employed multiple pathogenicity scores from the Single Nucleotide Polymorphism Database (dbSNP). RVs had to achieve a deleterious rating in at least 36 of 40 the dbSNP predictors (90%) to qualify as protein altering [41]. Conversely, we were more inclusive in our sample selection by retaining only unrelated persons of British ancestry, while Jurgens, *et al.* made no such exclusions (Figure 2.3) [41].

Theses methodological differences were reflected in *GCK* RV carrier rates. In our analysis, T2D was present in about 12% of carriers and 5% of non-carriers (Table 3.1). In contrast, they observed T2D in 48% of carriers and in 7% of non-carriers [41]. While more stringent variant filtering criteria and less exclusive sample selection may have contributed to a stronger signal for *GCK*, it is not clear if the same applied to all genes relevant to T2D. The balance between eliminating false negatives and retaining true positives is situational to the data available, so there are no universally agreed upon thresholds for RV MAF and pathogenicity scores [44]. Therefore, it may be necessary to include diverse parameters in RV analyses to enable comparisons between studies.

Other parameters to consider were the models on which the analysis was based, particularly our chosen genetic penetrance model and method of association: additive and allele burden, respectfully. In the additive penetrance model, the disease risk in each carrier is based on a continuous value determined by the sum of gene's alternate alleles [45]. On one hand, the additive model may be appropriate for T2D as it fits well into the common disease, common variant model [46]. On the other hand, the additive model may be inappropriate for RV burden analysis because it assumes too low of a penetrance for the protein altering mutations. Alternatively, different degrees of penetrance can be assumed by using the dominant or recessive models, where the disease risk in each carrier is based on a binary value. The carrier's disease risk is conditional on whether the variant genotype provides a sufficient alternate allele count: one/heterozygous in the dominant model and two/homozygous in the recessive model (Table 2.4) [45]. Across all cohorts, a majority of RV carriers had single heterozygous alternate genotypes in each gene, so adaption of the higher penetrance dominant model may not be too different from the current additive model. In contrast, use of the lower penetrance recessive model would then drastically reduce the number of carriers due to the rarity of homozygote alternate genotypes. The recessive carrier state can also be caused by compound heterozygotes, where two different heterozygous mutations appear on the same gene, but on different chromosomes, having the same effect as a homozygote [47]. Detection of compound heterozygotes requires phasing, where variants are assigned to the maternal and paternal chromosomes [47]. While phasing was not incorporated in the exome sequencing pipelines used in the UKB or DBGap, it is possible to phase variants using computational techniques *in-silico* [48].

Whichever model of genetic penetrance is used for RVs however, the allele burden test is not the only, or necessarily the best, option for association with T2D. While collapsing RVs by genes improves statistical power, this method implicitly makes two assumptions: uniform direction and magnitude of effect [49]. Within a given gene, different variants could have bidirectional effects, where some are deleterious and increase T2D risk while others are protective and decrease risk [49]. Some variants may have a large functional impact, such as affecting a protein's binding site, while others may result in a minor conformational change. The burden test does not account for any of these deviations from its assumptions, limiting its power to detect true associations [17]. In the UKB intracohort analysis for example, per unit increase in RV burden in the *ACE* gene conveyed 18% increased odds of risk for T2D. Although this finding did not reach the exome-wide significance threshold for multiple hypothesis testing, it is still somewhat alarming because it contradicts previous evidence that both LOF mutations in *ACE* and inhibition of angiotensin converting enzyme protects against T2D [22-24]. While it is possible that gain-of-function mutations were driving the RV gene burden association, in the model they would have assumed to be of equal magnitude to the more numerous LOF mutations and effectively been cancelled out [49].

Variants of bidirectional and varying magnitude of effect are handled better by another kind of association test that measure dispersion of variance in cases and controls [50]. One of these is the sequence kernel association test (SKAT), which may prove to be a versatile alternative to the allele burden test [49]. Multiple software is available that can run burden tests and SKAT in parallel, while also using SDA or other techniques to control for case-control imbalance [51].

Despite using the same T2D cases as the intracohort analyses, the inter-cohort comparisons to summary level data in GnomAD produced very divergent results. While it may seem promising that many cohorts had genes that achieved exome wide significance, the sheer abundance of these associations indicates a high false positive rate (Figure 3.2, Table S6-10). Some genes shared strong association signals across intracohort and inter-cohort analyses, suggesting that the associations for those gene were driven by enrichment in cases instead of depletion in controls. However, none of these replications qualified as independent, and it was impossible to tell how much systematic inflation affected the associations in GnomAD. This inflation was tempered by RV-EXCALIBER, as shown by the lower lambda values in all cohorts after implementation of individual (iCF) and gene correction (gCF) factors (Figure 3.2). The high genomic inflation pre-correction may have been caused by the compromise made to establish non-differential coverage (Section 2.5.2). Recall that binary alignment files for individual samples were not received, necessitating intersection between GnomAD and each cohort per-variant at a study-wide level (Figure 2.3). This procedure may have inadvertently restricted the number of variants in GnomAD, such that they were depleted relative to the number of variants in the cases [34]. It is possible that some groups of variants existed in linkage disequilibrium (LD) such that they were inherited together non-randomly [2]. A group of RVs that are in LD should not be counted as separate contributors to gene burden because they are less affected by selective pressures compared to RVs that are not in LD [3]. Failure to account for variants in LD can cause inflation of gene burden. While the GnomAD database was curated to filter out individuals with pediatric disease, its unlikely that all cases of T2D or other confounding adult-onset diseases were entirely removed. This phenotypic heterogeneity could have contributed to inflation, or otherwise added noise to any true association signals [52]. Despite the large size of GnomAD, the inter-cohort analyses was well-powered for fewer genes than the intracohort analyses, an average reduction of about 75% (Table 2.3, 2.5). This could have been caused by the gCF correction because its calibration of mutation load requires genes to be cross-referenced between cases and another cohort [7]. We used non-T2D controls for this step, though more permissive options may be available.

4.3.2 Improvements

There are several methodological options to pursue to improve statistical power of the RV associations. These include increasing the sample size, exploring alternative parameters and robust techniques for true associations, and better data preparation for RV-EXCALIBER.

A larger sample size would improve not only representation of the T2D trait, but also the diversity of the RVs potentially relevant to the disease. As previously mentioned, over 40K T2D cases and controls are available from exome sequencing studies on DBGap and the EGA (Table 1.4) [53]. In late 2021 or early 2022 the UKB is planning to release its remaining 300K exomes, which includes over additional 20K T2D cases [54]. This upcoming subset is large enough to elicit an independent analysis for its majority European ancestry, while the smaller number of non-European samples can be pooled together with corresponding DBGap and EGA cohorts. Cohort-specific kinship matrices can be incorporated as random effects to adjustment for familial allelic frequencies, permitting the inclusion of related individuals [41,55]. A technique normally reserved for common variation in GWAS, imputation of RVs is now possible with whole genome sequences from the Trans-Omics for Precision Medicine (TOPMed) Program [56]. Imputation is the prediction of initially uncalled variants using reference panels of thousands of haplotypes: groups of variants in LD that are inherited together [57]. Imputation would help fill in the gaps of exome coverage in each cohort, increasing the number of RVs analysed. The phasing required to detect compound heterozygotes, a genotype relevant to the recessive penetrance model, can be accomplished *in-silico* with information from haplotype reference panels [57]. Fittingly, the SHAPEIT2 software can be used to both impute and phase RVs in the exome sequencing data [48].

One aspect worth emulating from Flannick, *et al.*'s study was their multi-pronged approach, whereby several variant filtering thresholds and association methods were used [13]. In the methodology of this thesis, a modest addition of the parameters previously discussed would accumulate to 24 combinations: two MAF thresholds (0.01 and 0.001), two pathogenicity score criteria (MCAP and dbSNP predictors), three penetrance models (additive, dominant, recessive), and two association tests (burden and SKAT). While this worsens the multiple hypothesis problem, the increased versatility to detect true associations may justify the extra exploration. The stringency required for statistical significance would deepen, but the large sample sizes available may be sufficient.

There are a few changes to the methodology that may improve the comparison of T2D cases to GnomAD. Recent improvements to the application programming interface provided by the UKB now allow for sequential downloads of sample CRAM files. Coverage information can be extracted from each individually and saved at a fraction of the disk space, allowing for complete intersection with GnomAD by region at a per-sample basis. Along with imputation and phasing, haplotype reference panels can be used to find RVs in LD and prune them out to prevent overcounting [3]. While not including the most recent release of whole genome sequences, GnomAD should be updated to version 2.1 because it contains data of greatest relevance to this thesis. GnomAD v2.1 comes with a controls-only subset, which should reduce phenotypic heterogeneity for T2D [58]. Lastly, calibration of the gCF for RV-EXCALIBER could be done using rare synonymous variants in the T2D case samples instead of protein altering RVs in the control samples [7]. Since they are part of the same exome capture, synonymous variants should be under identical mutation sequencing bias as coding variants [7]. Furthermore, this would eliminate the need to cross-reference with the control cohort, preventing gene loss. While currently outsized by the UKB and the collective cohorts in DBGap and EGA, optimization of GnomAD as a summary level control is still worthwhile because it captures a superior range of ethnic diversity.

4.3.3 Future Resources

The following three resources and technologies have the potential to greatly improve the efficacy of discovering true RV associations with T2D. While contemporary, many require more time, development, or accessibility before they can be fully adopted for this research.

First, there are the plethora of exome sequencing data which, either in a partial or full capacity, are not available to the public. For research institutions or commercial entities that keep internal databases, the restricted access may be due to legal obligations, moral privacy concerns, financial incentives, or cost prohibitions [59]. Still, with enough effort and motivation, some of these challenges could be overcome so that more data is released into the world. Conversely, databases that are already available to the public may not share all relevant information. For example, DBGap does not appear to host binary alignment files for its exome sequencing studies. Another example are the disease-specific subsets provided by GnomAD, which has yet to include one for T2D-controls only [58]. While unknown if planned or not, a later inclusion of these features would be well appreciated.

Second, there are the steadily growing records of clinically relevant variants with hard supporting evidence, which are yet too small to fully replace computational predictive tools such as MCAP. Two prominent archives of these variants are ClinVar and HGMD, both publicly available, where the former is completely free, and the latter has a premium version requiring an annual subscription. Each has an impressive number of disease-associated variants, about 120K in Clinvar and 300K in the HGMD [37,38]. With millions of RVs in the UKB and other exome sequencing cohorts, Clinvar and HGMD currently have too few variants to use directly as pathogenicity filters without heavily compromising statistical power (Table 2.3). They have however contributed indirectly, such as how MCAP was constructed using HGMD RVs as training data (Jagadeesh et al., 2016). While empirical validation of all human variation would be very challenging, ClinVar and HGMD show its inevitability [60].

Third, whole genome sequencing (WGS) sequencing will ultimately replace exome sequencing with as the primary technology behind RV association, though this may not come true in the immediate future. The cost of WGS is decreasing, but it remains at least twice as expensive on average as exome sequencing [61]. The extra cost of WGS buys improvements in both common and rare variant analyses but may only be justified for the former. Without having to contend with exon capture issues, WGS does offer better coverage of coding regions compared to exome sequencing [62]. However, by far the largest areas with upgraded coverage in WGS are intergenic regions, which are more relevant to detection of common variants and identifying break points for copy number variations [63]. Considering that large, multi-thousand sample sizes are required for effective RV analyses, exome sequencing could remain more affordable than WGS for some time [17]. Though WGS does provide ancillary benefits to existing exome sequencing data, like imputation and phasing using WGS-derived haplotype reference panels from TOPMed [56]. While WGS will eventually become the gold standard for both common and rare variation, currently the latter is still best approached with exome sequencing.

4.4 Summary

The RV gene burden results showed clinical significance by implicating MODY genes in T2D, challenging the mutual exclusivity of the two diabetes subtypes, prompting follow up analysis of diagnostic biomarkers. The exome wide significant association of RV burden in the gene *PAM* has been corroborated by recent human cell line and common variant experimentation linking the copper metabolism gene to insulin secretion and T2D incidence. Further research into the effects of *PAM* RV burden on insulin secretion is warranted.

The ability of the analyses to capture true RV associations was limited by the insufficient size of the non-European cohorts, avoidable removal of related T2D cases, burden test inflexibilities, and too stringent variant exclusion in GnomAD intersection. Solutions include extracting non-Europeans from the UKB, allelic frequency adjustment in related individuals, SKAT as an alternative test, and redoing intersection with complete coverage. WGS and forthcoming genetic databases will supersede exome sequence-based analysis, but until then the best course of action is to improve current methodologies.

4.5 References

1. Naito, R., & Kasai, T. (2015). Coronary artery disease in type 2 diabetes mellitus: Recent treatment strategies and future perspectives. *World journal of cardiology*, *7*(3), 119–124. https://doi.org/10.4330/wjc.v7.i3.119

2. Slatkin M. (2008). Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. *Nature reviews. Genetics*, *9*(6), 477–485. https://doi.org/10.1038/nrg2361

3. Turkmen A, Lin S. (2017). Are rare variants really independent? *Genet Epidemiol*. 41(4):363-371. doi: 10.1002/gepi.22039.

4. Type 2 Diabetes Knowledge Portal. (2021). Providing data and tools to promote understanding and treatment of type 2 diabetes and its complications. Retrieved January 25 2020 from: https://t2d.hugeamp.org/

5. Sachdeva A, Cannon CP, Deedwania PC, Labresh KA, Smith SC Jr, Dai D, Hernandez A, Fonarow GC. (2009). Lipid levels in patients hospitalized with coronary artery disease: an analysis of 136,905 hospitalizations in Get With The Guidelines. *Am Heart J*. 157(1):111-117.e2. doi: 10.1016/j.ahj.2008.08.010.

6. Cho, Y. I., Mooney, M. P., & Cho, D. J. (2008). Hemorheological disorders in diabetes mellitus. *Journal of diabetes science and technology*, *2*(6), 1130–1138. https://doi.org/10.1177/193229680800200622

7. Lali, R., Chong, M., Omidi, A., Mohammadi-Shemirani, P., Le, A., & Paré, G. (2020). Calibrated rare variant genetic risk scores for complex disease prediction using large exome sequence repositories. *BioRxiv*, 2020.02.03.931519. https://doi.org/10.1101/2020.02.03.931519

8. McCulloch, D.W. (2019) Classification of diabetes mellitus and genetic diabetic syndromes. In: Mulder, J.E., Nathan, D.M., and Wolfsdorf, J.I. (Eds.), *UpToDate*. Retrieved March 17, 2021, from https://www.uptodate.com/contents/classification-of-diabetes-mellitus-and-genetic-diabetic-syndromes

9. Ovsyannikova, A. K., Rymar, O. D., Shakhtshneider, E. V., Klimontov, V. V., Koroleva, E. A., Myakina, N. E., & Voevoda, M. I. (2016). ABCC8-Related Maturity-Onset Diabetes of the Young (MODY12): Clinical Features and Treatment Perspective. *Diabetes therapy : research, treatment and education of diabetes and related disorders*, *7*(3), 591–600. https://doi.org/10.1007/s13300-016-0192-9

10. Borowiec, M., Liew, C. W., Thompson, R., Boonyasrisawat, W., Hu, J., Mlynarski, W. M., El Khattabi, I., Kim, S. H., Marselli, L., Rich, S. S., Krolewski, A. S., Bonner-Weir, S., Sharma, A., Sale, M., Mychaleckyj, J. C., Kulkarni, R. N., & Doria, A. (2009). Mutations at the BLK locus linked to maturity onset diabetes of the young and beta-cell dysfunction. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(34), 14460–14465. https://doi.org/10.1073/pnas.0906474106

11. Data-Field 2976: Age diabetes diagnosed (2021). *The UK Biobank*. Retrieved February 14 2021 from, https://biobank.ndph.ox.ac.uk/ukb/field.cgi?id=2976

12. Ahlqvist, E., Storm, P., Käräjämäki, A., Martinell, M., Dorkhan, M., Carlsson, A., Vikman, P., Prasad, R. B., Aly, D. M., Almgren, P., Wessman, Y., Shaat, N., Spégel, P., Mulder, H., Lindholm, E., Melander, O., Hansson, O., Malmqvist, U., Lernmark, Å., … Groop, L. (2018). Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. *The Lancet Diabetes & Endocrinology*, *6*(5), 361–369. https://doi.org/10.1016/S2213-8587(18)30051-2

13. Flannick, J., Mercader, J.M., Fuchsberger, C. *et al.* (2019). Exome sequencing of 20,791 cases of type 2 diabetes and 24,440 controls. *Nature* 570, 71–76. https://doi.org/10.1038/s41586-019-1231-2

14. Bousquet-Moore, D., Mains, R. E., & Eipper, B. A. (2010). Peptidylgycine α-amidating monooxygenase and copper: a gene-nutrient interaction critical to nervous system function. *Journal of neuroscience research*, *88*(12), 2535–2545. https://doi.org/10.1002/jnr.22404

15. Tümer, Z., Møller, L. (2010). Menkes disease. *Eur J Hum Genet*, 18, 511–518. https://doi.org/10.1038/ejhg.2009.187

16. Thomsen, S.K., Raimondo, A., Hastoy, B. *et al.* (2018). Type 2 diabetes risk alleles in *PAM* impact insulin release from human pancreatic β-cells. *Nat Genet* 50, 1122–1131. https://doi.org/10.1038/s41588-018-0173-1

17. Zhang, X., Basile, A., Pendergrass, S., & Ritchie, M. (2019). Real world scenarios in rare variant association analysis: the impact of imbalance and sample size on the power in silico. *BMC Bioinformatics*, *20*(46). doi: 10.1186/s12859-018-2591-6

18. Qin, H., Zhao, J. & Zhu, X. (2019). Identifying Rare Variant Associations in Admixed Populations. *Sci Rep*. 9, 5458. https://doi.org/10.1038/s41598-019-41845-3

19. GenomeAsia100K Consortium., Wall, J.D., Stawiski, E.W. *et al.* (2019). The GenomeAsia 100K Project enables genetic discoveries across Asia. *Nature*. 576, 106–111 (2019). https://doi.org/10.1038/s415a86-019-1793-z

20. Jones, A. G., & Hattersley, A. T. (2013). The clinical utility of C-peptide measurement in the care of patients with diabetes. *Diabetic medicine : a journal of the British Diabetic Association*, *30*(7), 803–817. https://doi.org/10.1111/dme.12159

21. Staten, M. A., Stern, M. P., Miller, W. G., Steffes, M. W., Campbell, S. E., & Insulin Standardization Workgroup (2010). Insulin assay standardization: leading to measures of insulin sensitivity and secretion for practical clinical care. *Diabetes care*, *33*(1), 205–206. https://doi.org/10.2337/dc09-1206

22. Gillespie, E. L., White, C. M., Kardas, M., Lindberg, M., & Coleman, C. I. (2005). The Impact of ACE Inhibitors or Angiotensin II Type 1 Receptor Blockers on the Development of New-Onset Type 2 Diabetes. *Diabetes Care*, *28*(9), 2261–2266. https://doi.org/10.2337/diacare.28.9.2261

23. Fountain JH, Lappin SL. (2020). Physiology, Renin Angiotensin System. *StatPearls* [Internet]. https://www.ncbi.nlm.nih.gov/books/NBK470410/

24. Pigeyre, M., Sjaarda, J., Chong, M., Hess, S., Bosch, J., Yusuf, S., Gerstein, H., & Paré, G. (2020). ACE and Type 2 Diabetes Risk: A Mendelian Randomization Study. *Diabetes Care*, *43*(4), 835 LP – 842. https://doi.org/10.2337/dc19-1973

25. Bayrak, C. S., Jain, A., Stein, D., Chaudhary, K., Nadkarni, G. N., Van Vleck, T., Puel, A., Boisson-Dupuis, S., Okada, S., Stenson, P. D., Cooper, D. N., Schlessinger, A., & Itan, Y. (2021). Identification of Discriminative Gene-level and Protein-level Features Associated with Gain-of-Function and Loss-of-Function Mutations. *BioRxiv*, 2021.01.01.424981. https://doi.org/10.1101/2021.01.01.424981

26. Katsonis, P., Koire, A., Wilson, S. J., Hsu, T. K., Lua, R. C., Wilkins, A. D., & Lichtarge, O. (2014). Single nucleotide variations: biological impact and theoretical interpretation. *Protein science : a publication of the Protein Society*, *23*(12), 1650–1666. https://doi.org/10.1002/pro.2552

27. Song, R. (2016). Mechanism of Metformin: A Tale of Two Sites. *Diabetes Care*, *39*(2), 187–189. https://doi.org/10.2337/dci15-0013

28. Khan M, Di Scipio M, Judge C, Perrot N, Chong M, Mao S, Di S, Nelson W & Paré G. A versatile, fast and unbiased method for estimation of Gene-by-environment

interaction effects on biobank-scale datasets. (2021). *Nature Genetics Technical Reports.* (In Preparation).

29. Naylor R, Knight Johnson A, del Gaudio D. (2018), Maturity-Onset Diabetes of the Young Overview. (Adam MP, Ardinger HH, Pagon RA, et al., editors). *GeneReviews* [Internet]. https://www.ncbi.nlm.nih.gov/books/NBK500456/

30. Category 221: NMR metabolomics QC indicators (2021). *The UK Biobank*. Retrieved February 17 2021 from, https://biobank.ctsu.ox.ac.uk/crystal/label.cgi?id=221

31. Boehnke, M., et al. (2021). Type 2 Diabetes Genetic Exploration by Next-generation sequencing in multi-Ethnic Samples. *T2D-GENES Consortium*. Retrieved March 19 2021 from, http://t2d-genes.sph.umich.edu/

32. UK Biobank launches one of the largest scientific studies. (2020). *The UK Biobank.* Retrieved March 25 2021, from https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/news/uk-biobank-launches-one-of-the-largest-scientific-studies

33. Lali, R., Chong, M., Omidi, A. et al. (2021). Calibrated rare variant genetic risk scores for complex disease prediction using large exome sequence repositories. *Nat Commun* 12, 5852. https://doi.org/10.1038/s41467-021-26114-0

34. Guo, M. H., Plummer, L., Chan, Y. M., Hirschhorn, J. N., & Lippincott, M. F. (2018). Burden Testing of Rare Variants Identified through Exome Sequencing via Publicly Available Control Data. *American journal of human genetics*, *103*(4), 522–534. https://doi.org/10.1016/j.ajhg.2018.08.016

35. Blumhagen, R. Z., Schwartz, D. A., Langefeld, C. D., & Fingerlin, T. E. (2020). Identification of Influential Variants in Significant Aggregate Rare Variant Tests. *BioRxiv*, 2020.10.01.322644. https://doi.org/10.1101/2020.10.01.322644

36. Buniello, A., MacArthur, J., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., Suveges, D., Vrousgou, O., Whetzel, P. L., Amode, R., Guillen, J. A., Riat, H. S., Trevanion, S. J., Hall, P., Junkins, H., Flicek, P., … Parkinson, H. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic acids research*, *47*(D1), D1005–D1012. https://doi.org/10.1093/nar/gky1120

37. Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., Karapetyan, K., Katz, K., Liu, C., Maddipatla, Z., Malheiro, A., McDaniel, K., Ovetsky, M., Riley, G., Zhou, G.,

Holmes, J. B., … Maglott, D. R. (2018). ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic acids research*, *46*(D1), D1062–D1067. https://doi.org/10.1093/nar/gkx1153

38. Stenson, P. D., Ball, E. V., Mort, M., Phillips, A. D., Shiel, J. A., Thomas, N. S. T., Abeysinghe, S., Krawczak, M., & Cooper, D. N. (2003). Human Gene Mutation Database (HGMD ®): 2003 update. *Human Mutation*, *21*(6), 577–581. https://doi.org/10.1002/humu.10212

39. Zhou, W., Zhao, Z., Nielsen, J. B., Fritsche, L. G., LeFaive, J., Gagliano Taliun, S. A., Bi, W., Gabrielsen, M. E., Daly, M. J., Neale, B. M., Hveem, K., Abecasis, G. R., Willer, C. J., & Lee, S. (2020). Scalable generalized linear mixed model for region-based association tests in large biobanks and cohorts. *BioRxiv*, 583278. https://doi.org/10.1101/583278

40. Zhou, W., Nielsen, J. B., Fritsche, L. G., Dey, R., Gabrielsen, M. E., Wolford, B. N., LeFaive, J., VandeHaar, P., Gagliano, S. A., Gifford, A., Bastarache, L. A., Wei, W. Q., Denny, J. C., Lin, M., Hveem, K., Kang, H. M., Abecasis, G. R., Willer, C. J., & Lee, S. (2018). Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature genetics*, *50*(9), 1335–1341. https://doi.org/10.1038/s41588-018-0184-y

41. Jurgens, S. J., Choi, S. H., Morrill, V. N., Chaffin, M., Pirruccello, J. P., Halford, J. L., Weng, L.-C., Nauffal, V., Roselli, C., Hall, A. W., Aragam, K. G., Lunetta, K. L., Lubitz, S. A., & Ellinor, P. T. (2020). Rare Genetic Variation Underlying Human Diseases and Traits: Results from 200,000 Individuals in the UK Biobank. *BioRxiv*, 2020.11.29.402495. https://doi.org/10.1101/2020.11.29.402495

42. Fardo, D. W., Charnigo, R., & Epstein, M. P. (2012). Families or Unrelated: The Evolving Debate in Genetic Association Studies. *Journal of biometrics & biostatistics*, *3*(4), e108.

43. Data-Field 21000: Ethnic Background (2021). *The UK Biobank*. Retrieved February 10 2021 from, https://biobank.ndph.ox.ac.uk/ukb/field.cgi?id=21000

44. Bomba, L., Walter, K. & Soranzo, N. (2017). The impact of rare and low-frequency genetic variants in common disease. *Genome Biol*. 18, 77. https://doi.org/10.1186/s13059-017-1212-4

45. Clarke, G. M., Anderson, C. A., Pettersson, F. H., Cardon, L. R., Morris, A. P., & Zondervan, K. T. (2011). Basic statistical analysis in genetic case-control studies. *Nature protocols*, *6*(2), 121–133. https://doi.org/10.1038/nprot.2010.182

46. Schork, N. J., Murray, S. S., Frazer, K. A., & Topol, E. J. (2009). Common vs. rare allele hypotheses for complex diseases. *Current opinion in genetics & development*, *19*(3), 212–219. https://doi.org/10.1016/j.gde.2009.04.010

47. Miller, D. B., & Piccolo, S. R. (2020). Compound Heterozygous Variants in Pediatric Cancers: A Systematic Review. *Frontiers in Genetics*, *11*. https://doi.org/10.3389/fgene.2020.00493

48. Delaneau, O., Marchini, J. (2014). The 1000 Genomes Project Consortium. *et al.* Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. *Nat Commun*, 5, 3934. https://doi.org/10.1038/ncomms4934

49. Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., & Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *American journal of human genetics*, *89*(1), 82–93. https://doi.org/10.1016/j.ajhg.2011.05.029

50. Neale, B. M., Rivas, M. A., Voight, B. F., Altshuler, D., Devlin, B., Orho-Melander, M., Kathiresan, S., Purcell, S. M., Roeder, K., & Daly, M. J. (2011). Testing for an unusual distribution of rare variants. *PLoS genetics*, *7*(3), e1001322. https://doi.org/10.1371/journal.pgen.1001322

51. Chen, MH., Pitsillides, A. & Yang, Q. (2021). An evaluation of approaches for rare variant association analyses of binary traits in related samples. *Sci Rep*. 11, 3145. https://doi.org/10.1038/s41598-021-82547-z

52. Manchia, M., Cullis, J., Turecki, G., Rouleau, G. A., Uher, R., & Alda, M. (2013). The Impact of Phenotypic and Genetic Heterogeneity on Results of Genome Wide Association Studies of Complex Diseases. *PLoS ONE, 8*(10), e76295. https://doi.org/10.1371/journal.pone.0076295

53. European Genome-Phenome Archive. (2021). EGA Consortium. Retrieve March 2 2021 from, https://ega-archive.org/

54. Exome Data Release FAQs. (2020). *The UK Biobank*. Retrieved February 18 2021 from https://www.ukbiobank.ac.uk/media/cfulxh52/uk-biobank-exome-release-faq_v9-december-2020.pdf

55. Goudet, J., Kay, T., & Weir, B. S. (2018). How to estimate kinship. *Molecular ecology*, *27*(20), 4121–4135. https://doi.org/10.1111/mec.14833

56. Kowalski, M. H., Qian, H., Hou, Z., Rosen, J. D., Tapia, A. L., Shan, Y., Jain, D., Argos, M., Arnett, D. K., Avery, C., Barnes, K. C., Becker, L. C., Bien, S. A., Bis,

J. C., Blangero, J., Boerwinkle, E., Bowden, D. W., Buyske, S., Cai, J., … Li, Y. (2019). Use of &gt;100,000 NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium whole genome sequences improves imputation quality and detection of rare variant associations in admixed African and Hispanic/Latino populations. *PLOS Genetics*, *15*(12), e1008500. https://doi.org/10.1371/journal.pgen.1008500

57. Shi, S., Yuan, N., Yang, M., Du, Z., Wang, J., Sheng, X., Wu, J., & Xiao, J. (2018). Comprehensive Assessment of Genotype Imputation Performance. *Human Heredity*, *83*(3), 107–116. https://doi.org/10.1159/000489758

58. Francioli, L., Tiao, G., Karczewski, K., Solomonson, M., & Watts, N. (2018). gnomAD v2.1. *Broad Institute*. Retrieved August 22 2020, from https://gnomad.broadinstitute.org/blog/2018-10-gnomad-v2-1/

59. Narayanasamy, S., Markina, V., Thorogood, A., Blazkova, A., Shabani, M., Knoppers, B. M., Prainsack, B., & Koesters, R. (2020). Genomic Sequencing Capacity, Data Retention, and Personal Access to Raw Data in Europe. *Frontiers in Genetics*, *11*. https://doi.org/10.3389/fgene.2020.00303

60. Stenson, P.D., Mort, M., Ball, E.V. *et al.* (2020). The Human Gene Mutation Database (HGMD®): optimizing its use in a clinical diagnostic or research setting. *Hum Genet*. 139, 1197–1207. https://doi.org/10.1007/s00439-020-02199-3

61. Schwarze, K., Buchanan, J., Taylor, J.C. *et al.* (2018). Are whole-exome and whole-genome sequencing approaches cost-effective? A systematic review of the literature. *Genet Med*. 20, 1122–1130. https://doi.org/10.1038/gim.2017.247

62. Belkadi, A., Bolze, A., Itan, Y., Cobat, A., Vincent, Q. B., Antipenko, A., Shang, L., Boisson, B., Casanova, J.-L., & Abel, L. (2015). Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proceedings of the National Academy of Sciences*, *112*(17), 5473 LP – 5478. https://doi.org/10.1073/pnas.1418631112

63. Gross, A.M., Ajay, S.S., Rajan, V. *et al.* (2019). Copy-number variants in clinical genome sequencing: deployment and interpretation for rare and undiagnosed disease. *Genet Med*. 21, 1121–1130. https://doi.org/10.1038/s41436-018-0295-y

64. Vergès B. (2015). Pathophysiology of diabetic dyslipidaemia: where are we?. *Diabetologia*, *58*(5), 886–899. https://doi.org/10.1007/s00125-015-3525-8

65. Nesto, R. W. (2008). LDL Cholesterol Lowering in Type 2 Diabetes: What Is the Optimum Approach? *Clinical Diabetes*, *26*(1), 8 LP – 13. https://doi.org/10.2337/diaclin.26.1.8

66. de Leeuw C, Mooij J, Heskes T,& Posthuma D. (2015). MAGMA: Generalized gene-set analysis of GWAS data. PLoS Comput Biol, *11*(4), e1004219. doi:10.1371/journal.pcbi.1004219

67. Bell, K. J. L., Hayen, A., Macaskill, P., Craig, J. C., Neal, B. C., Fox, K. M., Remme, W. J., Asselbergs, F. W., van Gilst, W. H., MacMahon, S., Remuzzi, G., Ruggenenti, P., Teo, K. K., & Irwig, L. (2010). Monitoring Initial Response to Angiotensin-Converting Enzyme Inhibitor–Based Regimens. *Hypertension*, *56*(3), 533–539. https://doi.org/10.1161/HYPERTENSIONAHA.110.152421

68. Johansson, B.B., Fjeld K., El Jellas K., Gravdal A., Dalva M., Tjora E., Ræder H., Kulkarni R.N., Johansson S., N.jølstad P.R., Molven A. (2018). The role of the carboxyl ester lipase (CEL) gene in pancreatic disease. *Pancreatology*. 18(1):12-19. doi: 10.1016/j.pan.2017.12.001.

69. Kleinberger, J. W., & Pollin, T. I. (2015). Undiagnosed MODY: Time for Action. *Current diabetes reports*, *15*(12), 110. https://doi.org/10.1007/s11892-015-0681-7

CHAPTER 5 – Conclusion

5.1 Objective

The contribution of common and low effect genetic variation on Type 2 Diabetes (T2D) risk has been extensively characterized, but this has not been enough to fully explain the disease's heritability [1]. Therefore, the overall goal of this thesis was to bridge this gap by uncovering the effects of rare and high effect variation on T2D risk. This was primarily assessed by rare variant (RV) gene burden: comparing the number of rare, pathogenic mutations in protein-coding genes between T2D cases and controls. Sample phenotype and exome sequencing data were downloaded from publicly available cohorts: the United Kingdom Biobank (UKB), the Korean Association Resource (KARE) project, the Metabolic Syndrome in Men Study (METSIM), and the San Antonio Mexican American Family Studies (SAMAFS). A secondary comparison of RV gene burden was made between T2D cases and the general population. The latter was approximated with summary level variant calls from The Genome Aggregation Database (GnomAD), while adjusting for inter-cohort differences using Rare Variant Exome CALIBration using External Repositories (RV-EXCALIBER) [2].

5.2 Findings and Follow-up

Exome wide significance with T2D risk was found with *GCK*, and suggestive association with *HNF4A*, which are causal for two of the three most common Mature Onset Diabetes of the Young (MODY) subtypes. Diagnosis of MODY is based on the mutual exclusion from both Type 1 Diabetes and T2D, making their biomarkers and risk factors good targets for subsequent RV burden analyses [3]. Another exome wide association was in the copper metabolism gene *PAM*, which recently has been implicated in the mechanism of insulin secretion [4]. *PAM's* effect on endogenous insulin should be tested in applicable cohorts. Results from the inter-cohort comparisons to GnomAD suffered from systematic inflation, which made it difficult to parse out true association signals from the noise. However, GnomAD could be made more viable as an external control by reducing phenotypic heterogeneity, increasing inter-cohort intersection coverage, and improving calibration of RV-EXCALIBER [2,6].

5.3 Limitations and Improvements

There were several limitations to the data and methods used throughout this thesis, though improvements are possible with available resources. METSIM, KARE, and SAMAFS were of insufficient sample size to effectively capture RV burden, which was further exacerbated because the ethnicities and ancestries of the latter two cohorts are underrepresented in genomics [7,8]. The samples needed for well powered, ethnically diverse analyses can accessed from DBGap, as well as subsets of studies on the European Database of Phenotypes and Genotypes and the UKB of non-European ancestry. While the majority European UKB yielded statistically significant results, the analysis may have been undercut by the cohort's unbalanced case-control ratio, loss of T2D cases due to the exclusion of related individuals, overcounting variants in linkage disequilibrium (LD), and false positive signals from lenient variant pathogenicity criteria [9,10]. While case-control imbalance, sample kinship, and variant LD can be addressed with methodological adjustments, determining the optimal variant pathogenicity criteria is complicated because overt stringency can incur false negative signals [11]. Furthermore, the additive penetrance model and RV burden association test are limited by their core assumptions on carrier definition and variant effect trajectory, respectfully [7]. At the cost of exacerbating the multiple hypothesis problem, alternative penetrance models and association tests can be employed for an overall more robust analysis. Finally, polygenic risk scores for T2D risk can be included as a covariate to account for possible effects of common genetic variation [10].

5.4 Closing Remarks

RV analysis will improve over time with growing databases of empirically determined clinically applicable variants and the ongoing publications of both exome and whole genome sequencing data [11,12]. However, current tools and resources are sufficient for a well powered study, as demonstrated in this thesis. The results presented interesting findings on RV in T2D, namely the prevalence of major MODY genes and exploring the role of *PAM* in insulin secretion. Evidently the gap in T2D heritability has not yet been fully bridged, but this thesis has helped establish a foothold.

5.5 References

1. Grotz, A.K., Gloyn, A.L. & Thomsen, S.K. (2017). Prioritising Causal Genes at Type 2 Diabetes Risk Loci. *Curr Diab Rep* 17, 76. https://doi.org/10.1007/s11892-017-0907-y

2. Lali, R., Chong, M., Omidi, A., Mohammadi-Shemirani, P., Le, A., & Paré, G. (2020). Calibrated rare variant genetic risk scores for complex disease prediction using large exome sequence repositories. *BioRxiv*, 2020.02.03.931519. https://doi.org/10.1101/2020.02.03.931519

3. McCulloch, D.W. (2019) Classification of diabetes mellitus and genetic diabetic syndromes. In: Mulder, J.E., Nathan, D.M., and Wolfsdorf, J.I. (Eds.), *UpToDate*. Retrieved March 17, 2021, from https://www.uptodate.com/contents/classification-of-diabetes-mellitus-and-genetic-diabetic-syndromes

4. Flannick, J., Mercader, J.M., Fuchsberger, C. *et al.* (2019). Exome sequencing of 20,791 cases of type 2 diabetes and 24,440 controls. *Nature* 570, 71–76. https://doi.org/10.1038/s41586-019-1231-2

5. Gillespie, E. L., White, C. M., Kardas, M., Lindberg, M., & Coleman, C. I. (2005). The Impact of ACE Inhibitors or Angiotensin II Type 1 Receptor Blockers on the Development of New-Onset Type 2 Diabetes. *Diabetes Care*, *28*(9), 2261–2266. https://doi.org/10.2337/diacare.28.9.2261

6. Francioli, L., Tiao, G., Karczewski, K., Solomonson, M., & Watts, N. (2018). gnomAD v2.1. *Broad Institute*. Retrieved August 22 2020, from https://gnomad.broadinstitute.org/blog/2018-10-gnomad-v2-1/

7. Zhang, X., Basile, A., Pendergrass, S., & Ritchie, M. (2019). Real world scenarios in rare variant association analysis: the impact of imbalance and sample size on the power in silico. *BMC Bioinformatics*, *20*(46). doi: 10.1186/s12859-018-2591-6

8. GenomeAsia100K Consortium., Wall, J.D., Stawiski, E.W. *et al.* (2019). The GenomeAsia 100K Project enables genetic discoveries across Asia. *Nature*. 576, 106–111 (2019). https://doi.org/10.1038/s41586-019-1793-z

9. Turkmen A, Lin S. (2017). Are rare variants really independent? *Genet Epidemiol*. 41(4):363-371. doi: 10.1002/gepi.22039.

10. Khan M, Di Scipio M, Judge C, Perrot N, Chong M, Mao S, Di S, Nelson W & Paré G. A versatile, fast and unbiased method for estimation of Gene-by-environment interaction effects on biobank-scale datasets. (2021). *Nature Genetics Technical Reports.* (In Preparation).

11. Lee, S., Abecasis, G. R., Boehnke, M., & Lin, X. (2014). Rare-variant association analysis: study designs and statistical tests. *American journal of human genetics*, *95*(1), 5–23. https://doi.org/10.1016/j.ajhg.2014.06.009

12. Belkadi, A., Bolze, A., Itan, Y., Cobat, A., Vincent, Q. B., Antipenko, A., Shang, L., Boisson, B., Casanova, J.-L., & Abel, L. (2015). Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proceedings of the National Academy of Sciences*, *112*(17), 5473 LP – 5478. https://doi.org/10.1073/pnas.1418631112