Discovery, fine mapping, and functional characterization of genetic susceptibility loci in type 1 diabetes

Catherine C. Robertson
Chapel Hill, North Carolina

BA, Connecticut College, 2009
MS, University of Michigan, 2014

A Dissertation presented to the Graduate Faculty of the University of Virginia in Candidacy for the Degree of Doctor of Philosophy

Department of Biochemistry and Molecular Genetics

University of Virginia
December, 2021

Abstract

Discovery, fine mapping, and functional characterization of genetic susceptibility loci in type 1 diabetes

Catherine C. Robertson, PhD in Biochemistry and Molecular Genetics, University of Virginia, 2021

Type 1 diabetes (T1D) is an autoimmune disease in which the immune system destroys the insulin-producing β cells of the pancreas, leading to elevated blood glucose. Twin and family studies suggest that about half of T1D risk is inherited. In Chapter 1, I describe the genetic etiology of T1D.

Historically, Northern European populations had the highest incidence of T1D, but rates are increasing in other groups. Genetic studies of T1D in African-ancestry populations have been limited in scope and size. In Chapter 2, analyzing genotypes from 3,949 African Americans, we find genetic associations with T1D risk are broadly concordant between African- and European-ancestry groups but also demonstrate the value of population-specific genetic risk prediction.

Genetic studies conducted over the past 45 years have identified about 60 loci associated with T1D risk, but causal variants are unknown in most regions. In Chapter 3, we analyze genotypes from 16,159 T1D cases, 25,386 controls and 6,143 trio families, to identify additional T1D-associated regions and, in 52 regions, define "credible sets" of variants most likely to be causal for T1D.

The vast majority of T1D credible variants are in non-coding regions of the genome, which obscures the causal genes and mechanisms underlying their association. In Chapter 4, we identify causal cell types, variants, and genes for T1D using chromatin accessibility profiling. We demonstrate strong enrichment of T1D credible variants in open chromatin from lymphocytes, and find five regions where T1D credible variants influence chromatin accessibility in $CD4^+$ T cells.

These analyses expand our understanding of the genetic basis of T1D. In Chapter 5, I discuss how continued work to understand genetic risk for T1D in diverse ancestry and later-onset populations will pave the way for effective precision medicine in T1D. High throughput approaches to test variant function in diverse cell contexts will accelerate the effort to definitively link causal genes to T1D etiology, providing novel therapeutic targets and guiding application of therapies.

Dedications

I would like to thank my PhD mentor, Stephen Rich, for the many scientific opportunities he has given me, his invaluable guidance for matters big and small, and his continued confidence in me. I would also like to thank Suna Onengut-Gumuscu, who has been a close mentor throughout my training, for her patience, support, and countless insights. Thank you to both Steve and Suna for exposing me to the collaborative and dynamic process of large-scale genetic and genomic studies.

Thank you to John Todd at the University of Oxford, who co-supervised the work presented in Chapters 3 and 4, for his continuous support and enthusiasm for the project. Thank you to Dan Crouch, Tony Cutler, and Jamie Inshaw, also from the University of Oxford, who have been excellent collaborators and from whom I learned a lot. Special thanks also to Linda Wicker for her numerous insights about the genetics of type 1 diabetes and immunology and her careful inspection and generous feedback on multiple drafts of the manuscript. Thank you to John Todd and Steve Rich for their wisdom to send Jamie Inshaw to Charlottesville for three weeks in 2017, which initiated a rewarding collaboration. Finally, thank you to Jamie Inshaw for his insights about statistical genetics, his extraordinary patience as we deliberated over every detail of our manuscript, and his continued friendship since finishing his PhD.

Thank you to my thesis committee, Stefan Bekiranov, Charles Farber, Mike Guertin, Coleen McNamara, Suna Onengut-Gumuscu, and Stephen Rich. Thank you to the wider UVA community, including faculty and peers who have contributed

to my education in biomedical science and academic research. In particular, thank you to: all the faculty who participated in the Core Course, which opened my eyes to experimental biology; Mike Guertin and his lab, who welcomed me and taught me about transcription factors; Charles Farber, Mete Civelek, and David Auble for their informal mentorship on multiple aspects of the PhD process; Janet Cross and Amy Bouton for their mentorship and leadership during my time as a BIMS student; and Wei-Min Chen for his advice on analytical and computational challenges. Thank you to the scientists at UVA responsible for generating the high quality genetic and genomic data analyzed in this thesis. In particular, thank you to Emily Farber in the Genome Science Laboratory and Becky Pickin in the Onengut-Gumuscu lab.

Thank you to the students, staff, and faculty of the Center for Public Health Genomics, for their friendship and community. Special thanks to CPHG trainees Basel Al-Barghouthi, Chani Hodonsky, and Olivia Sabik, for their camaraderie and guidance. Thank you to all the peers I worked with through the graduate student organizations Women in Medical Sciences (WIMS), the Graduate Biosciences Society (GBS), and Graduate Recruitment Initiative Team (GRIT), including (but not limited to) Tori Osinski, Erin Weddle, Brittany Martinez, Claire Ruddiman, Tiffany Wang, and Maya Cabot. Many thanks to Kristyna Kupkova and Yaseswini Neelamraju for their friendship and frequent lunch dates, and Ingrid Braenne for her friendship during her time at UVA and since returning to Germany.

Thank you to my mentors prior to my arrival at UVA, including and especially Matt Sampson, who mentored me at the University of Michigan and has continued helping me find my way since then. Also, thank you to Laura Scott and Mike Boehnke from the Department of Biostatistics at University of Michigan, who introduced me to the field of human genetics and supported me in pursuing a PhD in biomedical sciences. Thank you to Steve Campbell who invited me to do a funded postbac year of applied math at North Carolina State University, and to Spencer Muse who

introduced me to the concept of statistical genetics while I was there. And, finally, thank you to Christopher Hammond, my undergraduate advisor in mathematics at Connecticut College, who encouraged me to major in mathematics and supported me in applying to graduate school.

Thank you to my family. To my parents, Kim and Cary Robertson, for their unconditional support, and my siblings, Matt Robertson and Susie Lyons, for their humor and friendship. Thank you to my husband, Chris Dampier, for everything. And, while I giver her no credit for the completion of this dissertation, thank you to my daughter, Grace, for making me smile at least one hundred times a day.

Finally, I'd like to thank the many children and young adults living with type 1 diabetes, as well as their families and health care providers, who have chosen to participate in the studies described in this thesis. We are indebted to these generous individuals for everything we have learned about type 1 diabetes, knowledge which is currently guiding us towards life-changing and potentially curative treatments.

Attributions

The projects in this thesis represent collaborative efforts, involving numerous scientists, institutions, and studies. Additionally, much of the text is adapted from published works. Below, I outline major contributions from other scientists in each chapter, as well as, the relevant published works.

Chapter 2

This chapter is based on data and analyses described in the following publication:

> Onengut-Gumuscu S, Chen WM, **Robertson CC**, Bonnie JK, Farber E, Zhu Z, Oksenberg JR, Brant SR, Bridges SL, Edberg JC, Kimberly RP, Gregersen PK, Rewers MJ, Steck AK, Black MH, Dabelea D, Pihoker C, Atkinson MA, Wagenknecht LE, Divers J, Bell RA, SEARCH for Diabetes in Youth, Type 1 Diabetes Genetics Consortium, Erlich HA, Concannon P, and Rich SS. Type 1 diabetes risk in African-ancestry participants and utility of an ancestry-specific genetic risk score. *Diabetes Care*. 2019 Mar 1;42(3):406-15.

This study was conceptually designed by Suna Onengut-Gumuscu, Pat Concannon, and Stephen Rich. Collection and genotyping of samples for this study was led by my mentors Stephen Rich and Suna Onengut-Gumuscu, who obtained DNA samples from SEARCH for Diabetes in Youth study (SEARCH), Genetics of Kidneys in Diabetes (GoKinD), Barbara Davis Center (BDC), University of California, San Fran-

cisco (UCSF), and New York Control Population (NYCP). These DNA samples were managed and genotyped at the University of Virginia by Suna Onengut-Gumuscu and Emily Farber. Additional samples, from University of Alabama at Birmingham (UAB) and Consortium for the Longitudinal Evaluation of African Americans with Early Rheumatoid Arthritis (CLEAR), were genotyped at the Feinstein Institute for Medical Research under the supervision of Peter Gregersen. Genotype data for external validation of the genetic risk score were provided by Mark Atkinson and Pat Concannon from the University of Florida.

Genotype clusters and laboratory quality control were conducted by Suna Onengut-Gumuscu and Emily Farber. Subsequent quality control analyses were implemented by Wei-Min Chen and Jessica Bonnie. Association analyses outside the MHC were led by Wei-Min Chen, with contributions from myself and Jessica Bonnie. Genetic risk prediction analyses were implemented by Wei-Min Chen, with input from Suna Onengut-Gumuscu on SNP selection. Writing of the manuscript was led by Suna Onengut-Gumuscu, Wei-Min Chen, Henry Erlich, Pat Concannon, and Stephen Rich.

My contributions to this study included performing imputation, imputation validation, and association analysis of HLA alleles and amino acid sequences. In addition, I performed preliminary versions of quality control analysis and association analysis for regions outside the MHC on early versions of the data set and provided feedback and input on validation approaches for the African-specific genetic risk score. For the manuscript, I contributed text for the HLA analysis methods and results sections.

Chapters 3 and 4

These chapters are based on data and analyses described in the following publication:

> **Robertson CC**, Inshaw JRJ, Onengut-Gumuscu S, Chen WM, Flores Santa Cruz D, Yang H, Cutler AJ, Crouch DJM, Farber E, Bridges SL Jr., Edberg JC, Kimberly RP, Buckner JH, Deloukas P, Divers J, Dabelea

D, Lawrence JM, Marcovina S, Shah AS, Greenbaum CJ, Atkinson MA, Gregersen PK, Oksenberg JR, Pociot F, Rewers MJ, Steck AK, Dunger DB; Type 1 Diabetes Genetics Consortium; Wicker LS, Concannon P, Todd JA, and Rich SS. Fine-mapping, trans-ancestral and genomic analyses identify causal variants, cells, genes and drug targets for type 1 diabetes. *Nature Genetics*. 2021 Jun 14:1-0.

This study was conceptually designed by Pat Concannon, John Todd, and my mentor, Stephen Rich. Sample collection was organized by Stephen Rich, Suna Onengut-Gumuscu, and John Todd. David Dunger provided samples for genotyping through the GRID. Panos Deloukas provided Immunochip genotyping data through the UK Blood Service. Jane Buckner provided samples for genotyping and data from the BRI. S. Louis Bridges Jr provided samples for genotyping through the CLEAR consortium. Peter Gregersen provided samples for genotyping through the NYCP project. Jasmin Divers, Dana Dabelea, Jean Lawrence, Santica Marcovina and Amy Shah provided samples for genotyping and data through SEARCH. Carla Greenbaum and Mark Atkinson provided samples for genotyping through TrialNet. Robert Kimberly, Jeffrey Edberg, Marian Rewers, Andrea Steck, Jorge Oksenberg, and Flemming Pociot provided samples for genotyping through their affiliated institutions and research programs.

The vast majority of DNA samples were genotyped at UVA, where they were managed by Suna Onengut-Gumuscu and Emily Farber. Frozen T1DGC peripheral blood mononuclear cell samples for chromatin-accessibility profiling (ATAC-seq) were managed by Pat Concannon and Suna Onengut-Gumuscu. ATAC-seq data generation at UVA was supervised by Suna Onengut-Gumuscu and implemented by Becky Pickin. The generation of ATAC-seq data at the University of Oxford was led by Tony Cutler.

Wei-Min Chen and Suna Onengut-Gumuscu provided substantial analytical guid-

ance, particularly for genotype quality control analyses. Linda Wicker contributed to data interpretation, especially for fine-mapping and haplotype analyses. Dan Crouch provided statistical advice.

Nearly every analytical decision in this manuscript was deliberated over on trans-Atlantic conference calls between Jamie Inshaw and myself, with occasional guest appearances from Tony Cutler, Dan Crouch, Wei-Min Chen, Suna Onengut-Gumuscu, Linda Wicker, John Todd and Steve Rich. Jamie Inshaw implemented the case-control association analysis of unrelated subjects, meta-analysis of association results across analysis sub-cohorts, fine mapping of the European cohort with GUESSFM, fine mapping of the multi-ethnic cohort with PAINTOR, haplotype analyses of European cohort based on GUESSFM-prioritized credible sets, credible set enrichment analysis in open chromatin from diverse cell types, and the Priority Index drug target prioritization analysis. Jamie also generated the figures and investigated supporting published data for Section 3.3.5. I implemented raw genotype data processing and quality control, relationship inference and error correction in family data, genotype imputation to TOPMed and 1000 Genomes reference panels, assessment of imputation accuracy and coverage using whole genome sequencing, determination of ancestry groups and evaluation of sources of population stratification in case-control association analysis, family-based association analysis, annotating protein-altering T1D credible variants, processing and quality control for ATAC-seq data generated at UVA, generation of caQTL maps, and integration of caQTL maps with T1D association results. I generated the figures and investigated supporting published data for Section 4.3.3. David Flores processed and analyzed the ATAC-seq data generated at University of Oxford, with input and guidance from Tony Cutler. Hanzhi Yang performed the EMSA assays described in Section 4.3.3, under the supervision of Suna Onengut-Gumuscu. The manuscript was written by Jamie Inshaw and myself, under the supervision of John Todd and Stephen Rich, and with substantial input from Linda Wicker and

Suna Onengut-Gumuscu. Many other co-authors also provided useful feedback on the manuscript text and figures.

Contents

Chapter 1

Introduction

Sections of this chapter are adapted from:

> **Robertson CC** and Rich SS. Genetics of type 1 diabetes. *Current Opinion in Genetics & Development*. 2018 Jun 1;50:7-16.

1.1 Background on type 1 diabetes

1.1.1 Physiology and natural progression

Insulin is a peptide hormone that regulates glucose transport into muscle and adipose cells. Insulin is produced by β cells in the islets of the pancreas (Figure 1.1). In type 1 diabetes (T1D), previously referred to as "insulin-dependent diabetes," "autoimmune diabetes," and "juvenile diabetes," the immune system attacks and destroys the pancreatic β cells, causing permanent loss of insulin production (Figure 1.2). Without insulin, blood glucose levels become dangerously high, leading to serious and potentially fatal complications. Since the 1920's, exogenous insulin injections have been used to control blood glucose levels in T1D patients. Improved insulin therapy and delivery systems have substantially lessened the daily burden and long-term risks of living with T1D. However, worldwide, T1D continues to be associated with dimin-

ished quality of life and shortened life expectancy, particularly in communities with restricted healthcare resources.

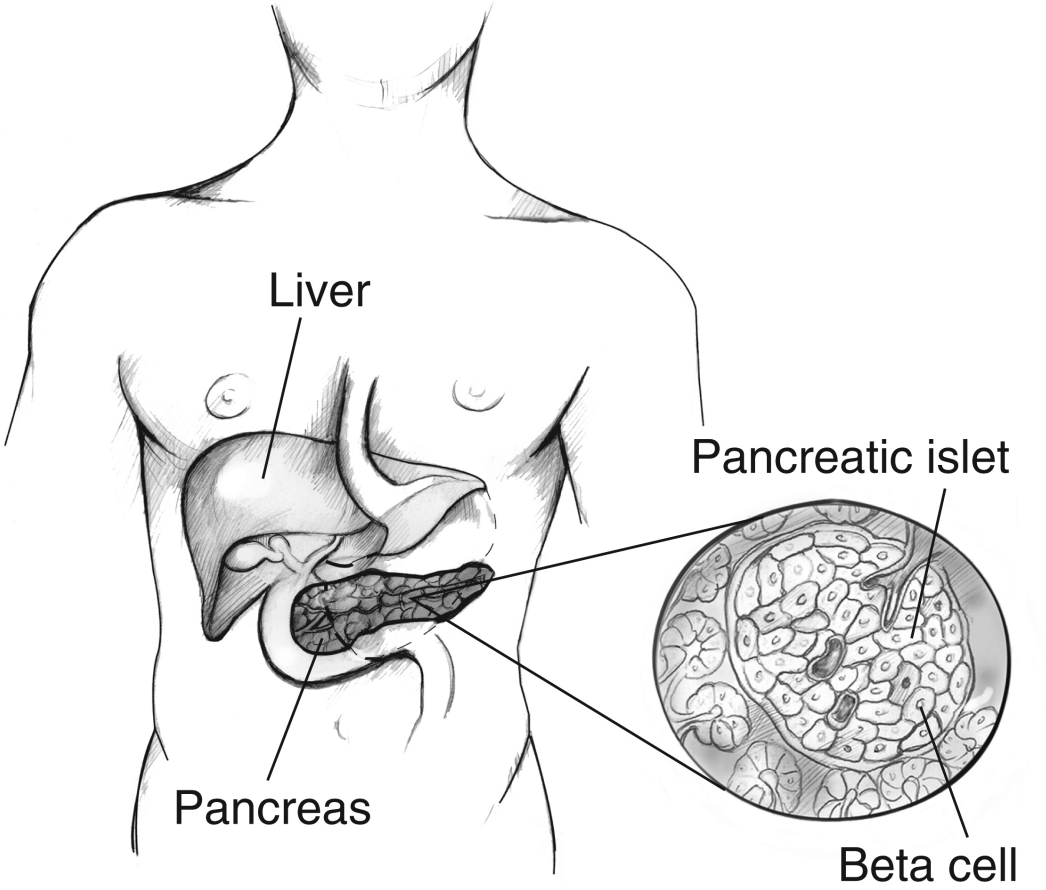

Figure 1.1: The pancreas is located in the abdomen behind the stomach. Islets within the pancreas contain β cells, which produce insulin. Image and caption obtained from the National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health. https://www.niddk.nih.gov/news/media-library/8034

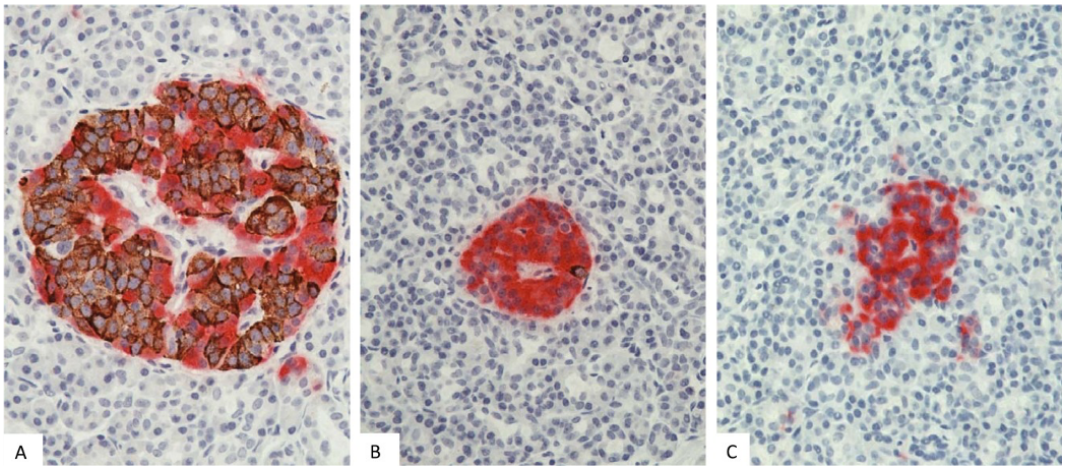

Figure 1.2: Islet architecture is altered during β-cell destruction. Individual islets from a single organ donor recently diagnosed with type 1 diabetes were immunostained for insulin (brown) or glucagon (red). Islets in the initial phase of peri-insulitis tend to have a normal architecture with both α and β cells present (A). Islets in which most of the β cells have been destroyed (B) are usually smaller and adopt a more condensed appearance. Occasionally, insulin-deficient islets adopt a more diffuse appearance (C), with loss of their typical cellular organization. Image and caption adapted from Morgan et al. 2014.

Based on available evidence, T1D is thought to develop through multiple stages (Insel et al. 2015). First, an unknown environmental trigger initiates a breakdown in immune tolerance to islet antigens in genetically susceptible individuals. Once initiated, autoimmunity against islet antigens causes progressive loss of β-cell function and mass during a pre-symptomatic phase lasting from months to years. Over time, injury to β cells begins to impair insulin production resulting in dysglycemia. Eventually, poor glucose control leads to overt symptoms and clinical diagnosis of diabetes. Continued injury to β cells causes complete lifelong dependence on exogenous insulin and potential complications to other organ systems. Pre-symptomatic islet autoimmunity can be detected by the presence of antibodies against specific islet antigens, including insulin autoantibodies (IAA), glutamic acid decarboxylase autoantibody (GADA), insulinoma-associated protein 2 autoantibody (IA-2A), and zinc transporter 8 autoantibody (ZnT8A). The presence of multiple islet autoanti-

bodies in an individual indicates a high risk of progression to T1D (Ziegler et al. 2013).

1.1.2 Prevalence and trends

Worldwide, about 1 in every 300 individuals are living with T1D. Although T1D can occur at any age (Thomas et al. 2018), it is easier to accurately diagnose in pediatric patients due to the lower prevalence of other forms of diabetes in this group. Thus, research on T1D has primarily focused on individuals developing symptoms in childhood or early adulthood. The incidence of T1D in children varies substantially across geographic regions (Figure 1.3). In Nordic countries (Finland, Sweden, and Norway), Saudia Arabia, and Kuwait, incidence rates exceed 30 in 100,000 children per year (Patterson et al. 2019). Meanwhile, some countries in Asia, South America, and Africa have estimated incidence rates of fewer than 2 in 100,000 per year (Patterson et al. 2019), suggesting more than a 15-fold difference in incidence rates between high- and low-incidence populations.

Incidence of T1D has increased worldwide in recent decades, primarily driven by steady increases in historically low-prevalence populations. In Europe between 1989 and 2003, Finland and Norway had the highest prevalence of T1D, yet the increases in annual incidence were lowest (2.7% in Finland and 1.3% in Norway) (Patterson et al. 2009). In contrast, Poland has a low T1D prevalence yet the incidence increased by 9.3% annually over the same period (Patterson et al. 2009). Meanwhile, in the United States from 2002 to 2012, the annual incidence of T1D increased overall, but differed by ancestry, with historically low-prevalence groups having higher annual rate of increase (4.7% in Hispanic and 2.2% in non-Hispanic black) compared to that in historically high-prevalence white youths (1.2% annual increase) (Mayer-Davis et al. 2017). These data point to a growing role for environmental exposures in the disease process, as well as, increased clinical recognition of T1D in historically low-prevalence

populations.

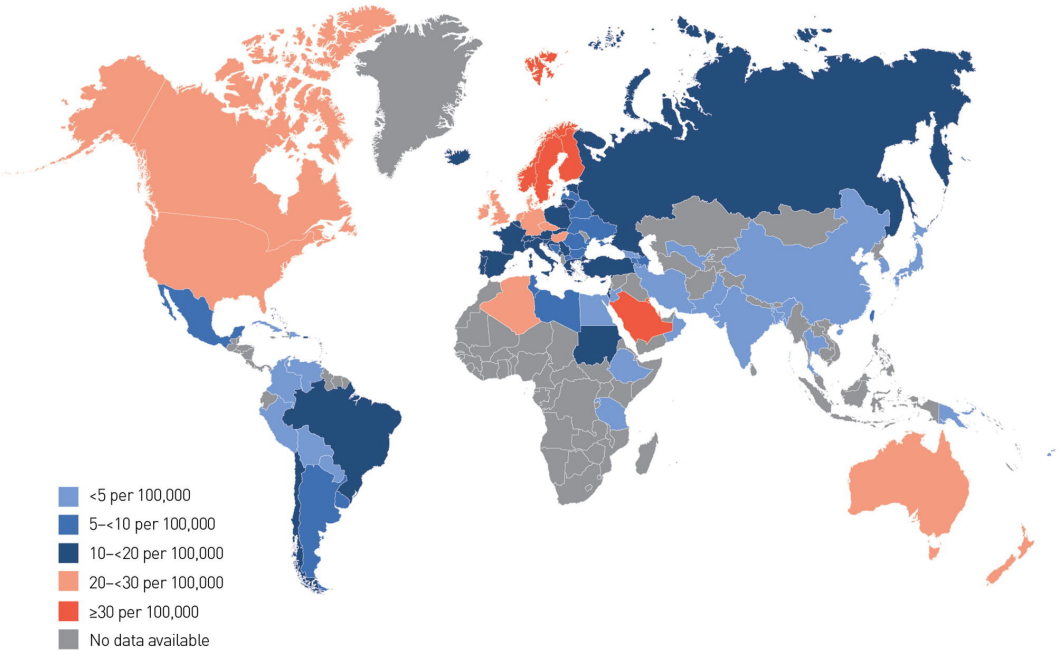

Figure 1.3: Map of age-sex standardised incidence rates (per 100,000) from publications of type 1 diabetes in children aged under 15 years. Image and caption from Patterson et al. 2019

1.1.3 Environmental triggers

Environmental factors may contribute to the initiation of islet autoimmunity or progression from islet autoimmunity to overt T1D in genetically susceptible individuals (Rewers and Ludvigsson 2016). Viral infection, particularly by enteroviruses, have the most support for a causal role in T1D development. Both epidemiological and animal studies implicate Coxsackie B enteroviruses in T1D (Richardson et al. 2009; Morgan and Richardson 2014; Stone et al. 2018). Prospective studies also show that respiratory infections early in life increase risk of islet autoimmunity among genetically high-risk children (Lönnrot et al. 2017).

One hypothesis proposes that physiological factors leading to increased insulin production can promote islet autoimmunity and progression to T1D through a "β-cell stress" mechanism (Roep et al. 2021). In particular, increased demands on β cells may

lead to post-translational modification of islet proteins involved in insulin secretion, creating immunogenic neoantigens that promote islet autoimmunity (McGinty et al. 2014; Delong et al. 2016). Potential β-cell stressors could include factors as diverse as rapid growth during puberty, viral infection, and psychological stress.

Many additional candidate triggers have been proposed, including vaccines, microbiome changes, increased sanitation in developed countries, and various dietary factors in early childhood (e.g., breastfeeding, cow's milk, and vitamin D), but, thus far, none are supported by convincing epidemiological evidence (Rewers and Ludvigsson 2016). Given existing findings, there is unlikely to be a single environmental trigger underlying T1D onset. Instead, several different exposures likely contribute to disease burden in the population, and, even in a single genetically susceptible individual, T1D may be triggered by combinatorial effects of these factors.

1.1.4 Treatment and prevention

Insulin was first isolated from dog pancreas secretions in 1921 (Banting et al. 1922). Within two years of this discovery, exogenous insulin injections were widely used to treat T1D in the United States and Europe (Bliss 1993). In the following decades, the pharmacodynamics of insulin therapies gradually improved (Bliss 1993), however, the immunogenicity of animal-derived insulin remained a challenge. In the 1970's, recombinant technology enabled synthesis of human insulin (Goeddel et al. 1979), which became the first genetically engineered therapy approved for clinical use in the United States (Keen et al. 1980).

More recently, advances in insulin delivery and glucose monitoring technologies, including "closed-loop systems" (Russell et al. 2014), have further improved quality of life for people living with T1D. Yet, even these tools cannot fully recapitulate the precise glucose control provided by functional β cells. For example, a recent trial demonstrated that, on average, children treated with a closed-loop insulin delivery

system still had glucose levels outside the target range more than 30% of the time (Breton et al. 2020). Due to the daily burden of glucose monitoring and long-term complications associated with poorly-controlled blood glucose, there is still a major role for novel approaches to treating this disease, particularly by replenishing functional β cells and preventing β-cell destruction by the immune system.

A number of immune-modulating therapies for T1D prevention or treatment have been explored in clinical trials with varying success (Skyler 2018). Recently, a 14-day course of teplizumab, an anti-CD3 monoclonal antibody that targets activated $CD8^+$ T cells, delayed T1D in high genetic-risk individuals by a median of two years (Herold et al. 2019). This success shows that appropriately timed immune-modulating therapy can alter the autoimmune process preceding T1D onset. Nonetheless, no treatment has been show to prevent T1D development, which motivates additional work to identify immune targets underlying the disease process in T1D patients.

1.2 Background on genetics

1.2.1 The genetic basis of human traits

The human genome contains about 3.4 billion nucleotides packaged into 24 chromosomes (1 through 22, X, and Y). This genome took shape in Africa about 200,000 years ago (McEvoy et al. 2011). Since then, humans have multiplied and migrated. The current genetic diversity of human beings across the globe has been shaped by a complex history involving many waves of migration and diverse evolutionary pressures. While the nucleotide sequence between any two humans is about 99.5% the same (1000 Genomes Project Consortium 2015), there are millions of locations in the genome where the nucleotide sequence might differ. These sites are called "genetic variants" or "polymorphisms," and the possible sequences at a given genetic variant are called "alleles."

Human genetics is the study of how genetic variants help to shape human traits. For some traits, the relationship to genetic variation is straightforward. These Mendelian traits, caused by genetic variation in one or a handful of genes, tend to follow simple inheritance patterns, often according the rules originally proposed by Gregor Mendel (Mendel 1865). For other "complex traits," including many common diseases, many genetic variants work together, often in concert with environmental factors, to shape disease susceptibility. The heritability of a trait is the proportion of phenotypic variance that can be explained by genotypic variance in a population. In other words, heritability is a measure of the relative contribution of genetic versus non-genetic factors to a trait. Heritability can range from 0 (no genetic influence) to 1 (completely determined by genetics). The heritability of a trait influences how well we can hope to predict the trait through genetics studies.

1.2.2 Measuring genetic variation in human populations

Genotyping is the practice of determining the alleles that a given individual carries at a particular genetic locus. For human autosomal chromosomes, a genotype will consist of a pair of alleles, one from each parental haplotype. Genotyping technology has evolved substantially in recent decades. Early approaches to systematically map human disease loci used restriction fragment length polymorphisms (RFLPs) to measure highly polymorphic markers across the genome (Botstein et al. 1980). RFLP experiments involve digesting DNA with restriction endonucleases and identifying alleles with gel electrophoresis. These linkage studies, which traced segregation of genetic markers and disease status in family pedigrees, were effective for identifying genetic causes of diseases that follow simple inheritance patterns (i.e., Mendelian genetic disorders). However, linkage analysis is not well-suited for identifying risk factors for genetically complex diseases due to the limited numbers of individuals and genetic markers that can be studied.

The Human Genome Project was an international effort to sequence all the euchromatic regions of the human genome, which was completed in 2004 (McPherson et al. 2001; Collins et al. 2004). The availability of a nearly complete map of the human genome, coupled with advances in DNA sequencing technologies ("next generation" or "high throughput" sequencing), made it possible to sequence the genomes of hundreds of individuals from diverse ancestral backgrounds. The International HapMap Project (The International HapMap Consortium 2003) and the 1000 Genomes Project (1000 Genomes Project Consortium 2015) provided a catalog of the most common genetic variants in human populations. With this catalog, genome-wide genotyping arrays were developed, providing an inexpensive way to systematically measure genetic variation across the genome (Figure 1.4, LaFramboise 2009). Genotyping arrays are particularly suited for measuring single nucleotide polymorphisms (SNP), genetic variants where the alleles differ by only a single nucleotide. Commercially available genotyping arrays typically genotype between 200 thousand and 2.5 million SNPs in a single experiment (Verlouw et al. 2021). Additional genetic variation can be studied by leveraging linkage disequilibrium to impute untyped variants using haplotype reference panels (see Section 3.1.2).

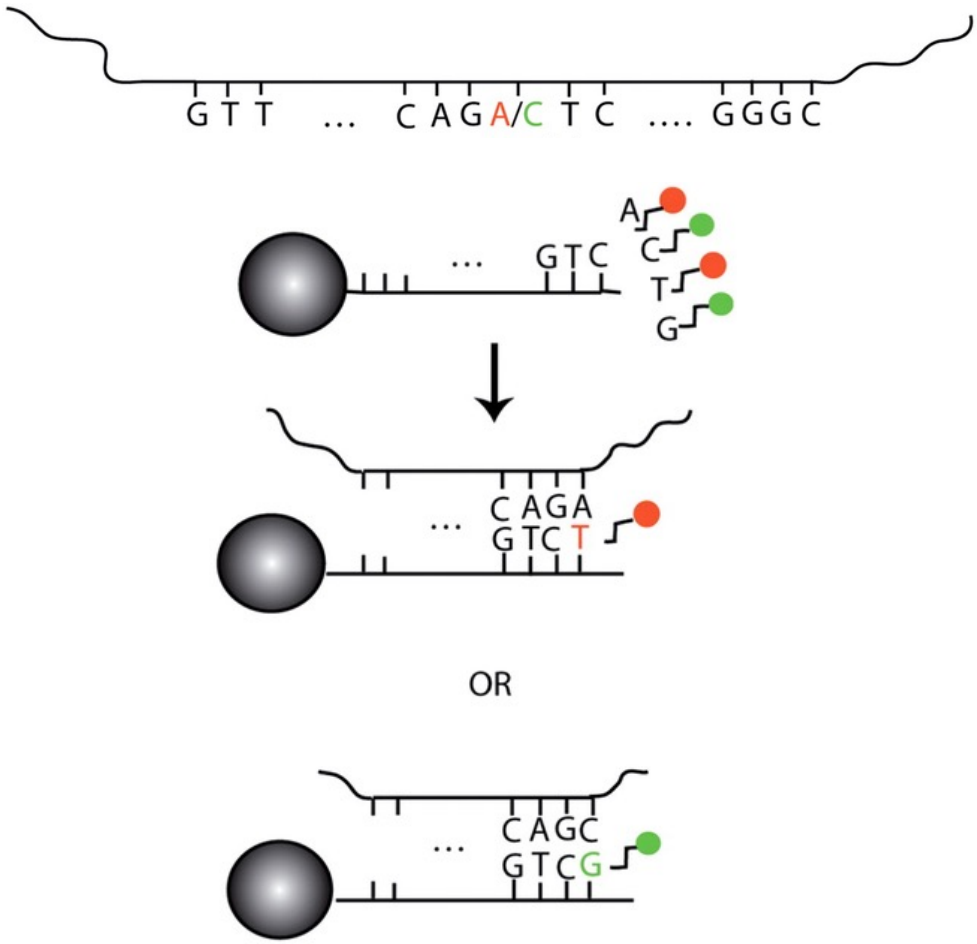

Figure 1.4: Overview of Illumina SNP array technology. At the top is the fragment of DNA harboring an A/C SNP to be interrogated by the probes shown. Attached to each Illumina bead is a 50-mer sequence complementary to the sequence adjacent to the SNP site. The single-base extension (T or G) that is complementary to the allele carried by the DNA (A or C, respectively) then binds and results in the appropriately-colored signal (red or green, respectively). Figure and caption adapted from LaFramboise 2009.

High throughput genotyping has enabled large-scale studies of disease cohorts, called genome-wide association studies (GWAS), and more recently population biobanks, to identify genetic factors underlying the full spectrum of human traits, including those with complex genetic architecture. Through many collaborative efforts to genotype tens of thousands of individuals, we now have robust maps linking over 55,000 unique genetic regions to almost 5,000 human traits (Figure 1.5), providing new in-

sights into disease etiology and avenues for therapeutic intervention (Loos 2020).

Figure 1.5: Published genome-wide significant ($p \leq 5 \times 10^{-8}$) associations with genetic variation on chromosome 1 over a ten-year period. Each vertical bar represents human chromosome 1. Each circle indicates association between a genomic region and a human trait or disease. Images obtained through the NHGRI-EBI GWAS Catalog.

1.2.3 The scale and distribution of human genetic variation

A typical human genome contains between 3.5 and 4.5 million non-reference alleles detectable by short-read whole genome sequencing, most of which are relatively common in human populations (1000 Genomes Project Consortium 2015). For example, more than 95% of variants within a typical human occur with an allele frequency greater than 0.5% globally (1000 Genomes Project Consortium 2015). Together, the global human populations represented in 1000 Genomes Phase 3 data (1000 Genomes Project Consortium 2015) contain about 8 million variants with allele frequency greater than 5%.

Rare variation is even more abundant. Somewhat paradoxically, while most of the genetic variation within a single individual is common, most of the variants observed in a population are rare (1000 Genomes Project Consortium 2015). Many more millions of rare variants have been identified through large sequencing projects. A recent sequencing study of 53,831 individuals genotyped using short-read whole genome sequencing identified about 400 million variants with allele frequencies less than 1% (Taliun et al. 2021). And while the number of known variants that are common in human populations will remain relatively stable, the total number of rare variants will likely continue to grow as the genomes of more individuals are sequenced.

1.2.4 Maps to mechanisms

The ultimate goal of human disease genetics is to improve understanding and treatment of human disease. While genetic maps are a valuable first step, the effect of genetic variants on disease risk must be interpreted in the context of molecular biology, cellular processes, and human physiology.

Genes as units of information

Originally proposed in 1958, the "central dogma of molecular biology" continues to provide a valuable framework for understanding the flow of information in biological systems (Crick 1970). Most importantly, DNA provides instructions in the form of genes, genes are transcribed into RNA, and RNA can be translated into proteins. The human genome contains about 20,000 protein-coding genes (Collins et al. 2004). Understanding the role of their protein products in diverse biological processes and diseases has been a major focus of molecular biology and human genetics research for decades. Thousands of genes have been interrogated experimentally and causally linked to human disease (Amberger et al. 2015, `OMIM.org`). Furthermore, comparative analysis of protein sequences and delineation of functional protein domains have allowed us to infer broad function for many human proteins through their evolutionary context (Lee, Redfern, and Orengo 2007). Nonetheless, the precise functions of most human genes in normal physiology remain unexplored.

Dynamic regulation of gene expression creates complex organisms

While protein-coding genes are essential units of biological information, they only make up about 1% of the human genome. The remaining 99% of the genome is thought to play a role in regulating expression of these genes across tissues, environments, and timepoints in human development. Deciphering the regulatory mechanisms of the non-coding genome remains a major challenge. Many assays using high throughput sequencing have been developed to measure molecular features involved in gene regulation, including transcription (Wang, Gerstein, and Snyder 2009), DNA-binding proteins (Park 2009), chromatin accessibility (Boyle et al. 2008; Schones et al. 2008; Buenrostro et al. 2013), and chromatin looping (Lieberman-Aiden et al. 2009). Together, these tools are starting to provide insights into mechanisms of gene regulation in complex organisms.

Genetic variation shapes complex traits through gene regulatory effects

Large-scale studies of complex human diseases have demonstrated that the majority of disease-associated variants are in non-coding regions of the genome (Maurano et al. 2012), making it difficult to infer causal genes. Thus, the value of large-scale genetic studies for understanding disease mechanism and providing pathways for therapeutic intervention will depend on our ability to decode the non-coding human genome. This realization has motivated collaborative efforts to systematically map genetic regulatory elements and variants, including the Encyclopedia of DNA Elements (ENCODE) (Moore et al. 2020), NIH Roadmap Epigenomics Mapping Consortium (Kundaje et al. 2015), Genotype-Tissue Expression (GTEx) project (GTEx Consortium 2020), and the International Common Disease Alliance (ICDA) (`www.icda.bio`). Through collective efforts like these, the field is beginning to delineate universal and cell-type specific regulatory elements and to explore how genetic variation may disrupt their activity. In some cases, molecular maps have led to testable hypotheses for causal variants and genes underlying complex disease associations (Musunuru et al. 2010; Small et al. 2011; Davis et al. 2019; Nasrallah et al. 2020). However, to date, only a small fraction of disease associations have known molecular mechanisms. Emerging technologies, for example high throughput genome editing and single cell sequencing (Fulco et al. 2016; Schraivogel et al. 2020; Pan et al. 2020; Morris et al. 2021), may help to accelerate the identification of causal variants and genes.

1.3 Genetic basis of T1D

Twin studies suggest a substantial genetic component to T1D risk. The reported percentage of monozygotic twins concordant for T1D diagnosis cross-sectionally has varied substantially, from 23% (Kaprio et al. 1992) to 54% (AH et al. 1981). However, a more recent study demonstrated that, in 65% of monozygotic twins initially

discordant for T1D, the unaffected individual will develop T1D by the age of 60 (Redondo et al. 2008). One study comparing monozygotic and dizygotic concordance rates in the Finnish national health system estimated that additive genetic effects account for 66% of T1D risk (Kaprio et al. 1992). This section describes our current understanding of the genetic factors contributing to T1D etiology.

1.3.1 Timeline of genetic discovery

Early studies identified associations of human leukocyte antigen (HLA) alleles (Singal and Blajchman 1973, Nerup et al. 1974) and variation in the insulin region (Bell, Horita, and Karam 1984) with "insulin-dependent diabetes" but not "non-insulin-dependent diabetes," which supported the emerging concept that these two conditions have distinct etiologies. As increasing evidence implicated autoimmunity as a defining feature of insulin-dependent diabetes (now called T1D) (Baekkeskov et al. 1982), diabetes diagnoses became more precise. Features such as presence of autoantibodies, age at onset, insulin resistance, and BMI were used to distinguish T1D from type 2 diabetes (T2D). With better-defined diagnostic criteria and an understanding that T1D is immune-mediated, three additional candidate genes with immune function were identified: *CTLA4* (Nisticò et al. 1996), *PTPN22* (Ladner et al. 2005), and *IL2RA* (Vella et al. 2005).

High throughput genotyping and increased sample sizes through aggregation of individual studies into consortia (e.g., the Type 1 Diabetes Genetics Consortium, T1DGC (Rich et al. 2006)) facilitated the identification of over 50 loci contributing to T1D risk (Smyth et al. 2006; Wellcome Trust Case Control Consortium 2007; Cooper et al. 2008; Barrett et al. 2009). In the 2009 T1DGC GWAS meta-analysis (Barrett et al. 2009), 41 distinct loci were associated with T1D risk. While these associations included known T1D loci, 27 were novel. Importantly, the majority of novel associations replicated in independent cohorts at genome-wide (18 out of 27)

or nominal (4 out of 27) significance levels.

In 2015, fine mapping of susceptibility loci with increased sample sizes refined our understanding of likely causal variants in T1D-associated regions (Onengut-Gumuscu et al. 2015). Allelic heterogeneity within a given locus was identified by conditioning on the most associated variant, discovering additional variants significantly associated with T1D risk near *IFIH1*, *IL2RA*, *INS*, *DEXI*, *PTPN2* and *TYK2*. Using imputation, four additional loci were identified in 2017 (Cooper et al. 2017).

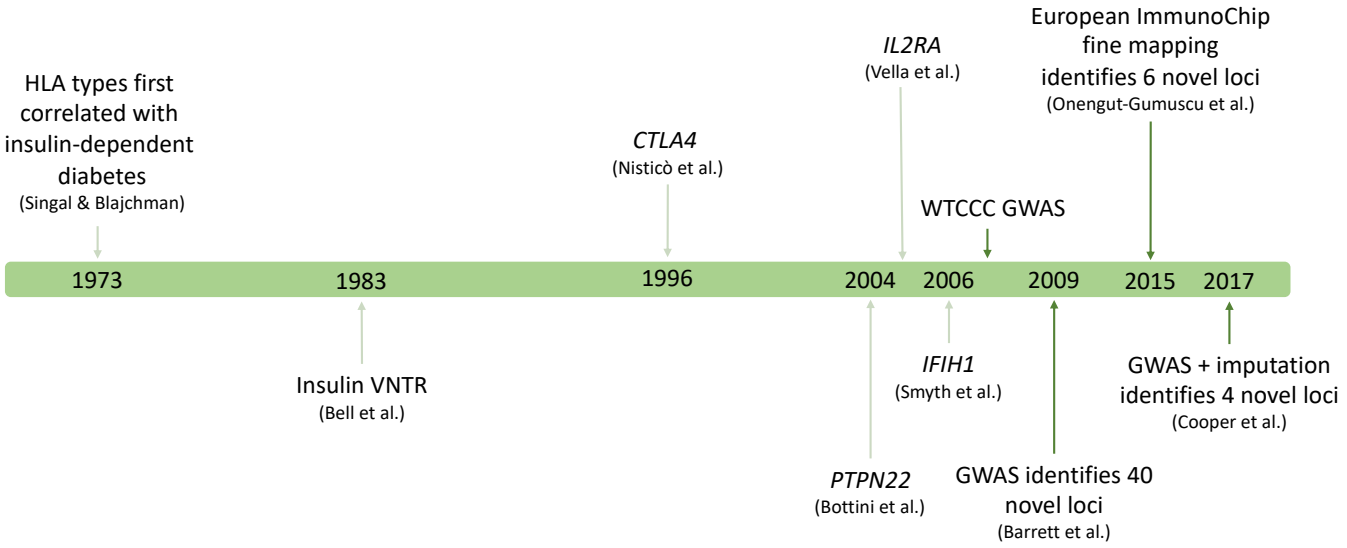

Figure 1.6: Timeline of genetic discovery in type 1 diabetes.

Together, these studies (Figure 1.6) provide a picture of the genetic architecture of T1D, where variation in HLA and insulin regions determine a large portion of disease susceptibility and the remainder of genetic risk is driven by small contributions from many loci (Figure 1.7). While many genetic risk factors underlying T1D susceptibility have been identified in European pediatric populations, the relative contribution of genes and exposures to T1D risk may vary across demographic groups and over

time. Expanding studies of T1D to a more representative patient population, including individuals with non-European ancestry and later disease onset, may reveal new disease genes and mechanisms.

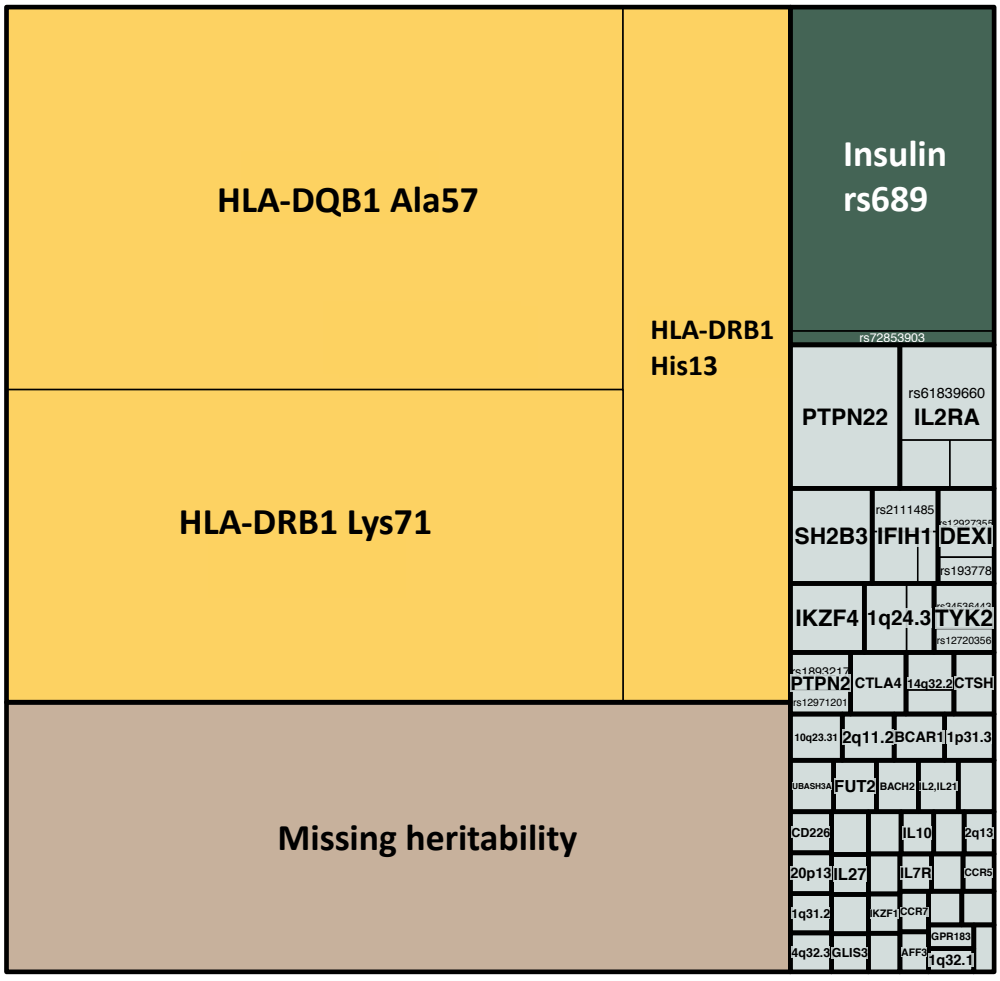

Figure 1.7: Visual representation of the genetic architecture of type 1 diabetes (T1D). The total area of the square represents the genetic basis of T1D, and the area of within each section is the proportion of heritability explained by each known T1D locus.

1.3.2 Disease genes and pathways

Across T1D-associated genetic loci, several genes are strongly suggested to play a role in disease etiology.

HLA genes

Variation in the major histocompatibility complex (MHC) accounts for the largest portion of genetic risk for T1D. However, due to the genetic complexity of the region, fully characterizing causal variants and mechanisms is challenging. The strongest T1D associations are with variants in the HLA class I (*HLA-A*, *-B*, and *-C*) and HLA class II (*HLA-DRB1*, *-DQA1*, *-DQB1*, *-DPA1*, and *-DPB1*) genes, which encode the MHC class I and II molecules. The presentation of peptide antigens by MHC molecules for recognition by T cells is an essential step in T-cell mediated adaptive immunity. HLA genes are among the most polymorphic genes in the human genome, with genetic variants concentrated in exons encoding the MHC molecule peptide binding groove. For further discussion of the biological importance and genetic structure of the MHC, see Chapter 2.

The association between HLA alleles and T1D risk was first recognized nearly 50 years ago (Singal and Blajchman 1973, Nerup et al. 1974). Subsequent work mapped the lead association to amino acid position 57 in *HLA-DQB1* (Todd, Bell, and McDevitt 1987), which affects the P9 peptide binding pocket of MHC class II HLA-DQ molecules. A mouse model that spontaneously develops autoimmune diabetes, the non-obese diabetes (NOD) mouse, also carries a distinct amino acid residue at position 57 of the *HLA-DQB1* murine homolog compared to other mouse strains (Todd, Bell, and McDevitt 1987).

Patterns of T1D-associated HLA class II alleles support the hypothesis that they confer risk or protection through structural changes in MHC molecules that alter peptide binding and presentation (Cucca et al. 2001). More recently, a bioinformatic approach relating HLA genetic variation with amino acid change and resulting protein structure showed that the majority of HLA-mediated association with T1D can be statistically explained by three amino acid changes that affect the binding pockets of MHC class II molecules: position 57 in *HLA-DQA1* lining the HLA-DQ P9 pocket,

and positions 13 and 71 in *HLA-DRB1* lining the HLA-DR P4 pocket (Hu et al. 2015). This confirmed the previous implication of the HLA-DQ P9 pocket and provided the first convincing evidence that the HLA-DR P4 pocket has a significant role in T1D, although it has been implicated in rheumatoid arthritis. In rheumatoid arthritis, the risk-conferring residues in the HLA-DR P4 pocket may be involved in binding of citrullinated peptides (Scally et al. 2013). A similar process may be involved with islet autoantigens, as suggested by recent work showing citrullination of β-cell proteins in response to inflammatory stress (McGinty et al. 2014; Babon et al. 2016; Buitinga et al. 2018).

Altered autoantigen presentation could influence risk of autoimmunity through multiple mechanisms. For example, a specific HLA class II allele, known to confer dominant protection from Goodpasture disease, binds a different register of the type $\alpha 3_{135-145}$ self-peptide, leading to increased abundance of regulatory T cells specific for this epitope and preserved immune tolerance to the endogenous type IV collagen protein (Ooi et al. 2017). Risk of islet autoimmunity may be shaped by a similar mechanism, where a small set of peptide-MHC complexes are only possible in the context of specific HLA-DQ P9 and HLA-DR P4 pockets. These peptide-MHC complexes may predispose or protect individuals from T1D through their effects on the relative abundance of regulatory versus conventional T cells specific to their epitopes. In addition to hypotheses involving structural effects of amino acid changes in MHC molecules on peptide binding, recent work has indicated potential regulatory effects for T1D-associated HLA class II alleles (Gutierrez-Arcelus et al. 2020; Fasolino et al. 2021), which may add to or modify the effect of coding sequence changes.

While MHC class II molecules likely play a causal role in T1D etiology, and perhaps explain a substantial portion of disease burden in pediatric disease, population studies suggest additional complexity underlying the MHC association with T1D. Early studies demonstrated genetic heterogeneity in the region (Rich, Weitkamp, and

Barbosa 1984), and more recent work with larger sample sizes confirm significant non-additive and interaction effects for HLA class II alleles (Hu et al. 2015). A sequencing approach to determine all HLA-DRB alleles suggested that *HLA-DRB3* alleles can modify the effects of *HLA-DRB1* alleles on T1D risk and are themselves associated with development of specific islet autoantibodies (Zhao et al. 2016). There is also evidence for additional MHC risk driven by variation beyond *HLA-DQB1* and *HLA-DRB1* alleles. Using additive models in pediatric-onset Northern European ancestry populations, genetic variation in HLA genes account for approximately 30% of T1D risk (Speed et al. 2012; Hu et al. 2015), of which only 80-90% was explained by HLA class II variation (Hu et al. 2015). This is consistent with work that has mapped T1D association to the MHC class I genes, *HLA-A* and *HLA-B* (Nejentsev et al. 2007), and non-HLA genes in the MHC region (Hippich et al. 2019). These studies underline the likelihood that MHC association with T1D risk is complex and mediated by multiple independent and interacting variants.

Insulin

The insulin gene (*INS*) represents the second most strongly associated locus with T1D risk. A highly variable polymorphism in the region was first identified using restriction fragment polymorphism analysis (Bell, Karam, and Rutter 1981). This polymorphic locus was subsequently shown to be a variable number tandem repeat (VNTR) (Bell, Selby, and Rutter 1982) at which individuals typically carry between 26 and 209 repeats of the sequence "ACAGGGGTGTGGGG." Individuals homozygous for the shorter class I VNTR alleles (26-63 repeats) are at increased risk of T1D, while individuals with one or two longer class III alleles (140-209 repeats) are substantially protected from T1D (relative risk $\approx$ 0.3) (Bell, Horita, and Karam 1984; Bennett et al. 1995).

The *INS* VNTR is located in the *INS* promoter. Protective class III VNTR alleles

correlated with higher insulin expression in the human thymus (Pugliese et al. 1997). Thus, one explanation for the dominant protective effect of class III VNTR alleles on T1D is that they promote negative selection of autoreactive T cells specific for insulin-derived peptides.

Insulin is an established T1D autoantigen, with many patients showing insulin autoantibodies (IAA) preceding disease onset. Consequently, a number of diverse mechanisms have been explored linking insulin mRNA expression, protein expression, and post-translational modification to T1D etiology. Notably, in humans, an alternative open reading frame within human *INS* mRNA has been shown to encode a highly immunogenic peptide that is targeted by $CD8^+$ T cells capable of killing β cells, and represents a pathway to islet autoimmunity and T1D (Kracht et al. 2017).

Negative regulators of T-cell activity

Several candidate causal T1D genes have been shown to affect T-cell signaling and activation pathways, including *PTPN22*, *CTLA4*, and *UBASH3A*. Genetic variation in each of these genes have been robustly associated with multiple autoimmune diseases, including T1D, suggesting potentially broad roles in shaping predisposition to T-cell mediated autoimmunity.

PTPN22 encodes the down-regulating protein LYP, which inhibits T-cell receptor (TCR) signal transduction by dephosphorylating tyrosine residues on two essential early signaling proteins, Lck and CD3. The minor allele of the common nonsynonymous coding SNP, rs2476601 (R620W), in *PTPN22* is strongly associated with T1D as well as other autoimmune diseases (Bottini et al. 2004). Functional studies of R620W indicate the presence of the W allele reduces response of both the T- and B-cell antigen receptors (Rieck et al. 2007). A second variant in *PTPN22* (rs56048322) was identified in those subjects with T1D who did not carry the R620W risk allele (Onengut-Gumuscu, Buckner, and Concannon 2006). Additionally, targeted se-

quencing in families with multiple T1D-affected individuals identified a cluster of rare deleterious variants associated with T1D risk, including novel frameshift mutations (rs869038523 and rs371865329) and a splicing variant rs56048322 (Ge et al. 2016). Functional studies of the splicing variant showed the novel isoform of LYP produced by the risk allele of rs56048322 resulted in reduced $CD4^+$ T-cell response to antigen stimulation. While the risk alleles at both SNPs, rs2476601 and rs56048322, exert similar effects on T-cell activation, rs56048322 has no known affect in B-cell receptor signaling, suggesting multiple actions of variation in *PTPN22* on T1D risk (Ge et al. 2016).

CTLA-4 is a cell surface receptor protein that plays a role in T-cell development and is a negative regulator of T-cell activation. CTLA-4 expression in both regulatory and effector T cells is essential for suppression of autoreactive T cell-mediated cytotoxicity (Wing et al. 2008; Ise et al. 2010). Genetic variants in the *CTLA4* gene have been associated with T1D as well as other autoimmune diseases, but the causal variants are not definitively known. One of the variants most associated with T1D, rs3087243, is strongly correlated with the length of an $(AT)_n$ dinucleotide repeat within the *CTLA4* 3' UTR (de Jong et al. 2016). Repeat length of this structural variant was shown to be inversely correlated with *CTLA4* mRNA expression in T-cell lines. A randomized control trial has demonstrated an effect of CTLA-4-targeted treatments to transiently reduce β-cell death in patients recently diagnosed with T1D (Orban et al. 2011).

UBASH3A encodes ubiquitin-associated and SH3 domain-containing A (UBASH3A, also referred to as STS2), which is expressed primarily in T cells and is a negative regulator of TCR signaling (San Luis et al. 2011). Two T1D risk variants (rs11202303 and rs80054410) are correlated with higher *UBASH3A* and lower *IL2* mRNA expression in human primary $CD4^+$ T cells upon TCR stimulation (Ge et al. 2017). UBASH3A has three functional domains (a ubiquitin-associated domain, SH3, and

a histidine phosphatase domain). While UBASH3A's role in down-regulating T cell signaling was originally attributed to phosphatase activity, it was recently shown to attenuate the NF-κB signaling pathway upon TCR stimulation via the UBA and SH3 domains (by a ubiquitin-dependent mechanism) (Ge et al. 2017). A follow-up study proposed a novel mechanism underlying *UBASH3A*-mediated regulation of T-cell activity where UBASH3A levels affect total abundance of TCR-CD3 complexes on the cell surface (Ge et al. 2019).

Interferon signaling genes

Two candidate causal genes for T1D have roles in interferon signaling and viral response. *IFIH1* encodes interferon induced with helicase C domain 1 (IFIH1), an innate immune receptor important for sensing viral infection. Both common and rare deleterious variants in *IFIH1* have been associated with T1D (Nejentsev et al. 2009). A recent study using human PBMCs and cell lines shows the common missense variant, rs1990760 (A946T), increases expression of type I interferons (Gorman et al. 2017). Additionally, knock-in mice for the risk variant had increased basal expression of type I interferons, improved survival from a lethal viral challenge, and increased susceptibility to streptozotocin-induced T1D (Gorman et al. 2017).

Tyrosine kinase 2 (TYK2) is a member of the JAK kinase family with roles in cytokine and type I interferon signaling via STAT phosphorylation. Multiple nonsynonymous variants in *TYK2* have a protective effect on T1D (Onengut-Gumuscu et al. 2015), potentially via down-regulation of interferon pathways (Marroqui et al. 2015). Inhibition of TYK2 in human pancreatic β cells attenuated double-stranded RNA-induced apoptosis by reducing type I interferon signaling and MHC class I protein expression (Marroqui et al. 2015).

Mechanisms implicated by noncoding regions

As with other complex traits, the majority of T1D-associated variants lie in noncoding regions with putative roles in regulating gene expression (see Section 1.2.4). In recent fine-mapping of T1D-associated loci (Onengut-Gumuscu et al. 2015), credible variants were enriched in regions containing enhancers active in thymus, $CD4^+$ and $CD8^+$ T cells, $CD19^+$ B cells and $CD34^+$ stem cells (Onengut-Gumuscu et al. 2015). These findings suggest that a sizable portion of T1D-associated variants likely exert their effects on disease susceptibility within one or more of these cell types. However, it does not rule out the possibility that some loci may act in other cell types, including within the pancreatic islets.

1.3.3 Genetic susceptibility across disease stages

Genetic risk factors are thought to contribute to different stages of T1D development (Figure 1.8) (Table 1.1). Genetic variants associated with initial autoantibody positivity may differ from those associated with eventual T1D diagnosis. Several studies have been designed to characterize the progression of children at risk of T1D to islet autoimmunity and clinical disease, including the Finnish Pediatrics Register (FPDR) (Ilonen et al. 2017), The Environmental Determinants of Diabetes in the Young (TEDDY) (Krischer et al. 2019), and the Diabetes Autoimmunity Study in the Young (DAISY) (Steck et al. 2011). Evidence emerging from these and other studies are consistent with diverse genetic and environmental factors driving initiation of islet autoimmunity and progression to T1D.

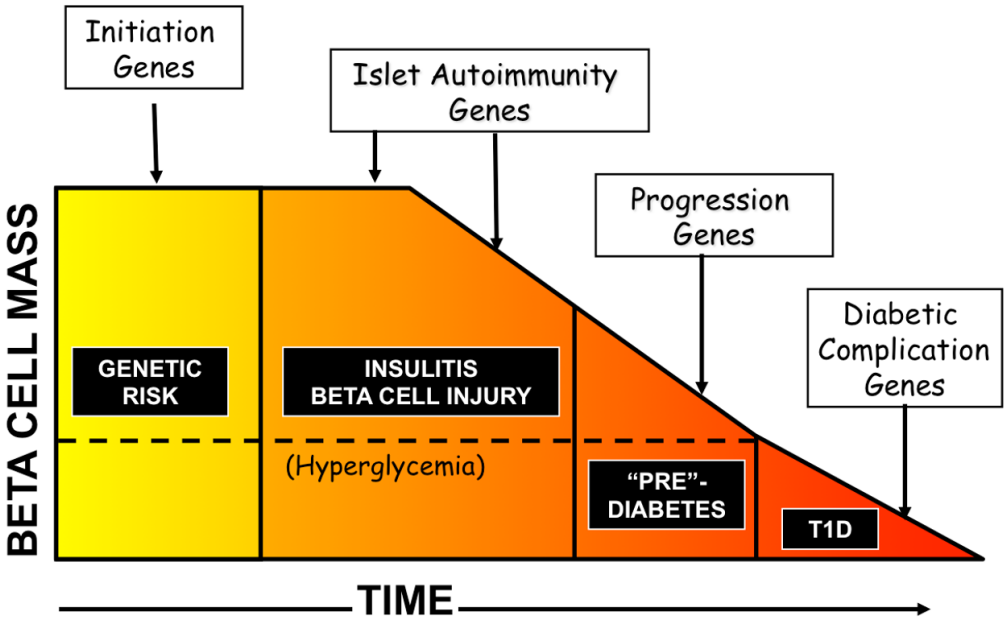

Figure 1.8: Stages in development of type 1 diabetes. A model proposes that disease is caused by immune-mediated destruction of insulin-secreting β cells in the pancreas, with genetic factors implicated at each stage. Image adapted from Atkinson and Eisenbarth 2001.

Early stages and emergence of islet autoimmunity

Pre-symptomatic islet autoimmunity can be detected by the presence of antibodies against specific islet antigens. Typically, insulin autoantibodies (IAA) or glutamic acid decarboxylase autoantibodies (GADA) appear first, and additional autoantibodies are acquired over time through antigen spreading (Krischer et al. 2015). By the time of T1D diagnosis, most patients have two or more islet autoantibodies (Ilonen et al. 2017).

However, T1D autoantibody positivity is dynamic, and the number and combination of autoantibodies in a patient may change over time. Autoantibodies can also be lost between seroconversion and progression to diabetes. Among subjects with IAA

as a primary autoantibody (i.e., the first autoantibody observed), IAA was no longer present at diagnosis 25% of the time (Ilonen et al. 2017).

The primary autoantibody detected at seroconversion may be indicative of the self-antigen that initiated autoimmunity. Genetic variation in HLA class II genes have been associated with which primary autoantibody is observed at seroconversion (Krischer et al. 2015; Ilonen et al. 2016) (Table 1.1). Additionally, patterns of autoantibodies at the time of diagnosis are correlated with primary autoantibodies at seroconversion (Ilonen et al. 2017). The ability to infer primary autoantigens based on combinations of autoantibodies observed at T1D diagnosis could be useful for inferring T1D subgroups in both research and clinical settings.

Table 1.1: Genetic associations with varying stages and features of T1D development. DIPP, Diabetes Prediction and Prevention; FPDR, Finnish Pediatric Diabetes Register.

Publication	Study	Stage or phenotype	Regions and genes implicated
Törn et al. 2015	TEDDY	Islet autoimmunity	1p13.2 (*PTPN22*), 11p15.5 (*INS*), 12q13.2 (*IKZF4*), and 12q24.12 (*SH2B3*)
Brorsson et al. 2015	T1DGC	Presence of GADA	3q28 (*LPP*)
Brorsson et al. 2015	T1DGC	Presence of IA-2A	1q23 (*FCRL3*) and 11q13 (*RELA*)
Brorsson et al. 2015	T1DGC	Presence of gastric parietal cell antibodies	2q24.2 (*IFIH1*)
Krischer et al. 2015	TEDDY	IAA as primary antibody	HLA DR4-DQ8
Krischer et al. 2015	TEDDY	GADA as primary antibody	HLA DR3-DQ2
Ilonen et al. 2016	DIPP and FPDR	Primary autoantibody	HLA DR-DQ haplotypes
Ilonen et al. 2017	DIPP and FPDR	Primary autoantibody	11p15.5 (*INS*), 12q13.2 (*IKZF4*)
Steck et al. 2017a	TrialNet	Time to development of multiple autoantibodies	2q33.2 (*CTLA4*)
Steck et al. 2017a	TrialNet	Progression to T1D from time of seroconversion	HLA DR, 9q24.2 (*GLIS3*), and 12q24.12 (*SH2B3*)
Ilonen et al. 2016	DIPP	Progression to T1D from time of second antibody	no associatons with HLA DR
Krischer et al. 2017	TEDDY	Progression to T1D from time of second antibody	6q23.3 (*TNFAIP3*) and 11p15.5 (*INS*)
Törn et al. 2016	TEDDY	T1D among DR4/4 subjects	HLA *C3*
Onengut-Gumuscu et al. 2020	DAISY	Time from first antibody to T1D onset	20p12.1 (*TASP1*), 1q21.3 (*MRPS21-PRPF3*), 2p25.2 (*NRIR*), and 3q22.1 (*COL6A6*)

Even before islet autoantibodies are observed, anergic autoreactive T and B cells specific for islet autoantigens can be found in patients (Smith et al. 2015). Thus, it is possible that mechanisms governing initial autoreactive repertoires (i.e., escape of central tolerance) are separate from those that ultimately break peripheral immune tolerance. Further examination of genetic and environmental factors underlying emergence of anergic islet antigen-specific T and B cells in healthy individuals may provide insights into the earliest stages of disease.

Secondary antigens and progression to T1D

Rate of progression from islet autoimmunity to T1D is correlated with the number of islet autoantibodies. Individuals who progress to multiple autoantibody positivity are more likely to be diagnosed with diabetes within 10 years of seroconversion (Steck et al. 2016; Bingley, Boulware, and Krischer 2016). High serum levels of IAA are also associated with increased risk of progression to T1D (Steck et al. 2016). These data suggest that increased efficiency of epitope spreading and increased autoreactive cell activity is correlated with faster progression of disease.

Factors contributing to epitope spreading and progression to T1D may be distinct from those initiating autoimmunity. While specific HLA DR-DQ genotypes and haplotypes have been established as risk factors for the development of islet autoimmunity and are predictive of primary autoantibodies, studies examining the role of HLA and non-HLA genotypes in progression to clinical disease have produced heterogeneous and sometimes conflicting results (Table 1.1). The lack of replication of genetic associations with rate of progression to T1D may reflect differences in the subject selection, environmental exposure and small sample size. Well-powered follow-up studies will be needed to clarify the role of HLA and other T1D-associated risk variants with progression from islet autoimmunity to T1D.

1.3.4 Early versus late-onset disease

T1D etiology may differ for early onset versus later onset disease. In histological analysis of pancreas tissue from individuals with recent-onset T1D, diagnosis at a younger age correlated with fewer insulin-containing islets and a higher proportion of $CD20^+$ B cells infiltrating inflamed islets (Arif et al. 2014; Leete et al. 2016). Although based on limited samples due to the difficulty of obtaining pancreas tissue from recent-onset individuals, these results suggests that children with early onset T1D (e.g., < 7 years) may have a more aggressive form of disease. Genetic studies of T1D also indicate that disease mechanisms may differ by age of onset. T1D-associated HLA risk alleles are also associated with earlier onset of disease (Valdes et al. 2012), and the effect of HLA alleles appears to be more potent in childhood than later in life (Inshaw et al. 2020). Outside the HLA, genetic variation also has a stronger effect on risk of early onset T1D (< 7 years) than older onset T1D (Inshaw et al. 2017; Inshaw et al. 2020). In subjects with a milder phenotype (i.e., single islet autoantibody and less severe dysglycemia at onset of T1D), variants in a T2D-associated locus (*TCF7L2*) were associated with T1D (Redondo et al. 2017). Relatively few genetic studies have focused on identifying genetic risk factors for late-onset T1D. Modest associations with late-onset ($>$ 14 yrs) T1D were observed with KIR haplotypes (Traherne et al. 2016). Expanding studies of later onset T1D may reveal new disease genes and mechanisms.

1.4 Applications of T1D genetics

Early efforts to predict T1D risk relied solely on high-risk HLA genotypes, which account for about 40% of the genetic risk (Barker et al. 2008). Inclusion of recently refined non-HLA risk variants and more powerful predictive modeling approaches have improved risk prediction (Sharp et al. 2019). As the sensitivity and specificity

of T1D risk prediction increases, practical and novel applications of genetic risk scores are emerging.

1.4.1 Population screening

While the positive predictive value of T1D genetic risk scores is likely to remain low due to low disease prevalence, their high sensitivity and specificity will make them useful for population screening. The Fr1da study pioneered population screening for T1D in Bavaria, Germany (Raab et al. 2016). In this initiative, children between the ages of 2 and 5 years are screened for islet autoantibodies by their primary care pediatricians to detect early islet autoimmunity. A potentially more cost-effective strategy would involve an initial T1D genetic risk assessment to narrow the population for follow-up autoantibody testing, although there would be subjects at low-to-moderate genetic risk who may still develop autoantibodies. Effective population screening and disease monitoring tools will facilitate enrollment of high-risk individuals in clinical trials of immune modulation during pre-symptomatic disease (Rich 2017).

1.4.2 Diagnosis in low prevalence groups

An important consideration when diagnosing T1D is distinguishing it from other forms of diabetes (Figure 1.9), which is more difficult in groups where T1D is lower prevalence or T2D is higher prevalence, including adults, overweight, and non-European groups. Recently, models of T1D genetic susceptibility have been used to assist differential diagnosis of T1D, T2D, and monogenic diabetes (Oram et al. 2016; Patel et al. 2016).

A recent analysis of UK Biobank participants demonstrated that genetic risk scores can be used to characterize disease prevalence in large cross-sectional populations (Thomas et al. 2018). A previously-defined T1D genetic risk score based on 29 common variants was applied to 379,511 unrelated individuals of European

descent. All subjects were ranked according to their T1D genetic risk score, and risk groups were defined by T1D risk scores above the median (high-risk) or below the median (low-risk). Since the T1D risk score was not associated with T2D risk, the expectation is that the frequency of T2D would be the same in the T1D risk groups. The number of individuals with T1D was genetically defined as the excess of individuals with diabetes among the high-risk group compared to the low-risk group. Using this approach, the age of onset of T1D was evenly distributed across the first six decades of life. Additionally, among the genetically defined T1D cases, there were no differences in clinical characteristics or risk of diabetic ketoacidosis between age groups (before or after 30 years) (Thomas et al. 2018).

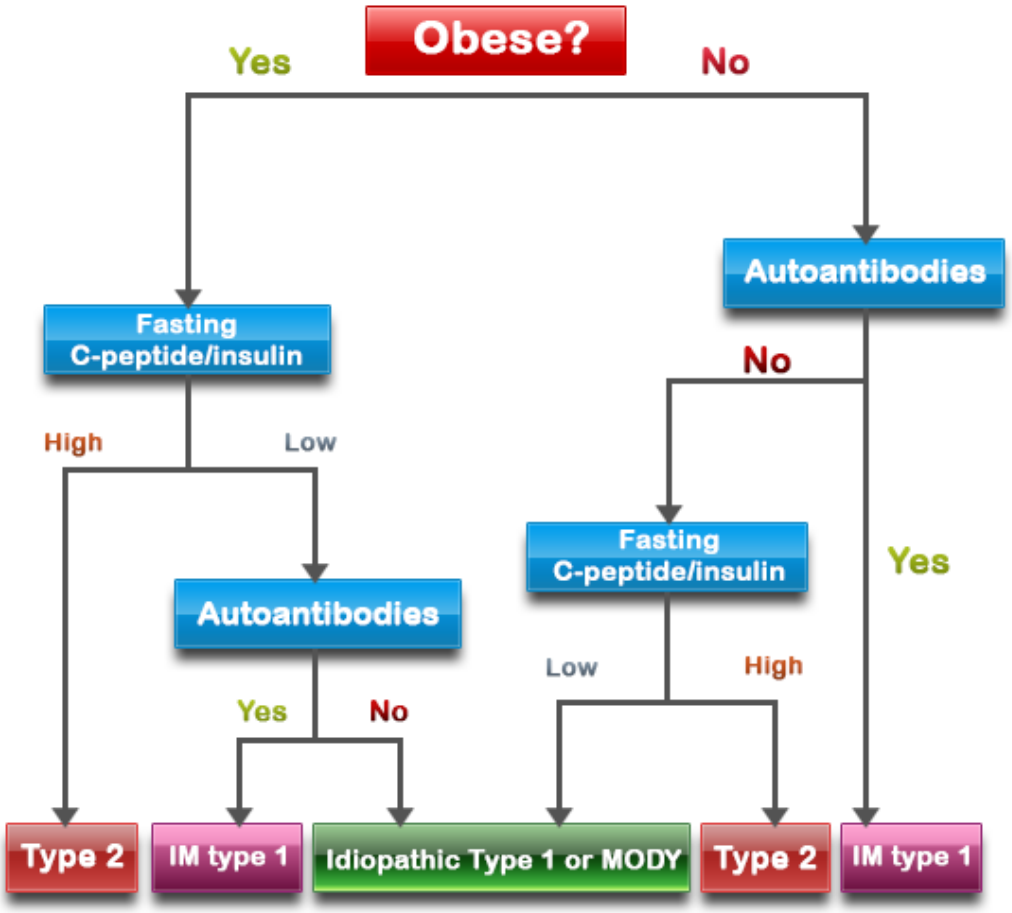

Figure 1.9: Flowchart showing tests to distinguish between different types of diabetes in adults. Obtained from https://www.diabetes.co.uk/which-type-of-diabetes.html

1.5 Motivation for this thesis

Our understanding of the genetic basis of T1D has only recently moved beyond the HLA genes and a few strong immune-based candidates. With increased sample size and genomic technologies, the catalogue of T1D genetic associations has grown to over 50 loci.

Genetic risk prediction will play an important role in facilitating early intervention in T1D by identifying high risk individuals. Furthermore, knowledge of genetic risk

of T1D has the potential to aid in the identification of the non-genetic risk factors for T1D. For example, the stratification of subjects into high genetic risk and low genetic risk could be used in a longitudinal cohort study, under the assumption that those subjects who develop islet autoimmunity and progress to T1D despite having a low genetic risk score would have been exposed to an unusually high burden of environmental risk factors.

However, numerous gaps in knowledge remain. The genetic information gathered to date is dominated by pediatric onset in European-ancestry populations. Genetic evaluation of non-European populations by whole genome sequencing will likely identify novel risk genes and novel variants in known genes. Additionally, genomic interrogation at all stages of the etiologic process in T1D should identify novel genetic factors that can be used for prediction of those at risk of developing islet autoantibodies as well as those who will have fast or slow progression to disease.

Successful application of genetics to T1D will depend on effective prevention of β-cell death using biologics and immunosuppressants. Precisely defining the molecular characteristics of genetic variation associated with disease risk, initiation, or progression is a promising approach to guide targeted therapies.

This thesis uses genetics to expand our understanding of why some people develop T1D, while most do not. In Chapter 2, we use the first large-scale study of T1D genetics in African-ancestry individuals to explore African-specific HLA alleles associated with T1D. In Chapter 3, we identify novel T1D risk regions and use optimized statistical methods to delineate potential causal variants underlying these associations. In Chapter 4, we use functional genomic data to generate hypotheses about causal mechanisms driving T1D associations. Together, these analyses help to clarify the variants, genes, and mechanisms that form the genetic basis of T1D susceptibility. In Chapter 5, I will discuss how these findings will guide future research and application of T1D genetics.

Chapter 2

HLA association with type 1 diabetes in an African American cohort

Sections of this chapter are adapted from:

> Onengut-Gumuscu S, Chen WM, **Robertson CC**, Bonnie JK, Farber E, Zhu Z, Oksenberg JR, Brant SR, Bridges SL, Edberg JC, Kimberly RP, Gregersen PK, Rewers MJ, Steck AK, Black MH, Dabelea D, Pihoker C, Atkinson MA, Wagenknecht LE, Divers J, Bell RA, SEARCH for Diabetes in Youth, Type 1 Diabetes Genetics Consortium, Erlich HA, Concannon P, and Rich SS. Type 1 diabetes risk in African-ancestry participants and utility of an ancestry-specific genetic risk score. *Diabetes Care*. 2019 Mar 1;42(3):406-15.

2.1 Background

2.1.1 The adaptive immune system

The immune system is a collection of cells and organs that protect organisms from disease caused by infection, cancer, or physical injury. In vertebrates, the immune system is broadly divided into two domains: innate and adaptive. Innate immune cells recognize non-specific patterns (e.g., molecules shared by many pathogens) and respond immediately (e.g., by engulfing the pathogen upon detection). Meanwhile, adaptive immune cells respond to highly specific antigens. On the first encounter with a pathogen, the adaptive immune response is substantially delayed compared to the innate response. However, adaptive immune cells can remember this first encounter, such that upon subsequent infection they are able to eliminate the pathogen very rapidly. This memory mechanism allows an organism to acquire "immunity" to many common pathogens over its lifetime.

The specific adaptive immune cells that facilitate immunological memory are called B lymphocytes and T lymphocytes (also referred to as B cells and T cells). B and T lymphocytes are activated through cell surface receptors that recognize highly specific antigens, the B-cell receptor (BCR) and T-cell receptor (TCR), respectively. Both the BCR and TCR are encoded by genes in germline DNA. However, during lymphocyte development, a process called V(D)J recombination (Figure 2.1) scrambles the DNA sequence in regions of these genes such that each developing cell contains a unique receptor. As a result, a single human contains B and T cells expressing tens of millions of unique BCR and TCR sequences. The extraordinary diversity of BCR and TCR sequences ensures that every human contains adaptive immune cells capable of recognizing, responding to, and remembering the unique molecular signature of nearly any pathogen.

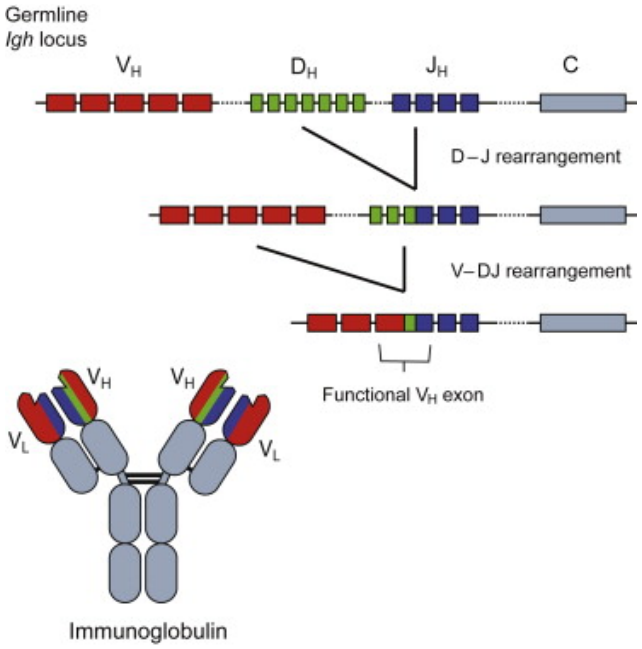

Figure 2.1: Schematic of V(D)J recombination in the immunoglobulin heavy chain locus, which encodes the B-cell receptor (BCR). Figure and caption adapted from Little et al. 2015.

2.1.2 The major histocompatibility complex

While BCRs on B cells can recognize diverse antigens directly in the extracellular environment, TCRs on T cells can only recognize proteins that have been partially degraded and presented on the surface of other host cells. The major histocompatibility complex (MHC) is a region of vertebrate genomes encoding the cell surface proteins, called MHC molecules, which are responsible for presenting antigen peptides for recognition by TCRs. In humans, the genes encoding MHC molecules are referred to as human leukocyte antigen (HLA) genes and are clustered together in the MHC region on the short arm of chromosome 6 (Figure 2.2).

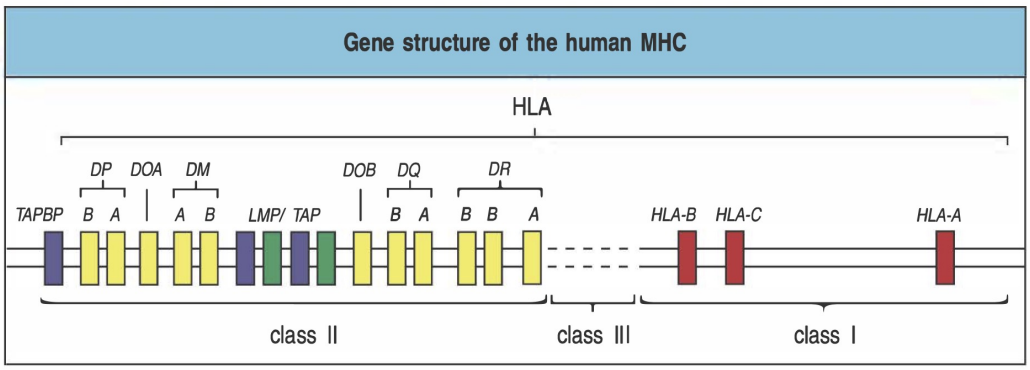

Figure 2.2: Genetic structure of human genes encoding MHC class I and II molecules. Figure adapted from Murphy, Travers, and Walport 2012.

There are two primary classes of MHC molecules. MHC class I molecules are expressed by all nucleated cells and present peptides to $CD8^+$ T cells. MHC class II molecules are expressed exclusively by professional antigen presenting cells (APCs) and present peptides to $CD4^+$ T cells. Classically, MHC class I molecules present peptides derived from cytosolic antigens, while MHC class II molecules present peptides from internalized extracellular proteins (Figure 2.3).

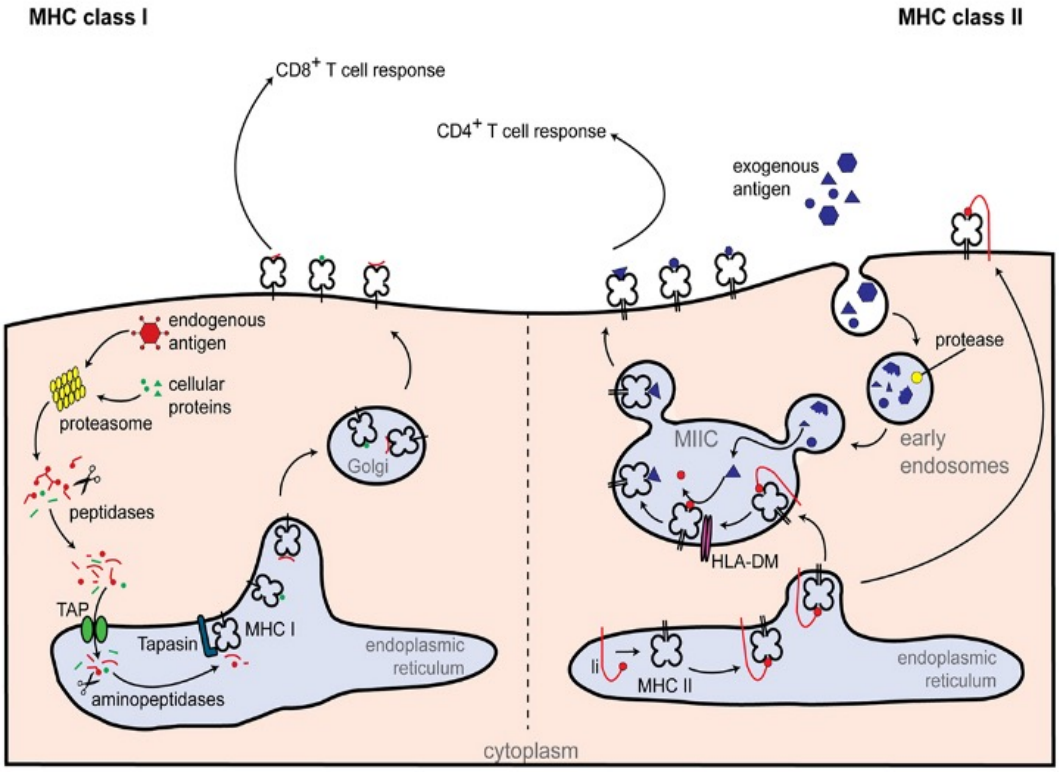

Figure 2.3: Classical routes of antigen presentation by MHC class I and II molecules. Class I antigen presentation (left): Proteasomes generate peptides from all proteins present within the cell. Peptide fragments are transported to the endoplasmic reticulum and loaded onto the MHC class I molecule. MHC class I-loaded complexes are transported to the cell surface. Class II antigen presentation (right): extracellular antigens are taken up by APCs. Phagosomes fuse with lysosomes, and proteolytic enzymes cleave the proteins into small peptides. MHC class II molecules from the endoplasmic reticulum are delivered to the phagolysosomes and loaded with peptide. Peptide-loaded MHC class II complexes are transported to the cell surface. Figure and caption adapted from Neerincx et al. 2013.

2.1.3 MHC diversity

The two MHC molecules have distinct structural features, but both are heterodimers encoded by unique α- and β-chains and both contain a peptide-binding groove (Figure 2.4). The amino acid sequence lining the peptide-binding groove influences which antigen peptides are capable of binding the MHC molecule for presentation on the cell surface. Thus, the repertoire of peptides available for recognition by T lymphocytes

depends on the combinations of α- and β-chain sequences available to form MHC class I and II molecules.

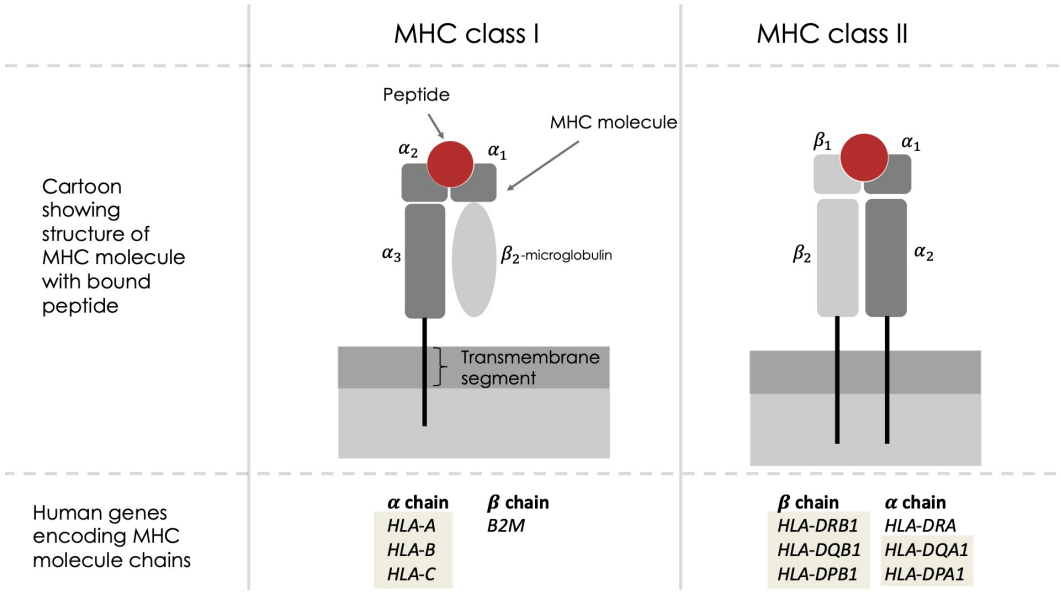

Figure 2.4: The general structure of MHC class I and II molecules (top) and the human genes that encode them (bottom). Highly polymorphic HLA genes are highlighted in yellow.

Two features of the MHC locus ensure that there is substantial diversity in MHC molecule peptide-binding groove sequences both within individuals and across the human population.

MHC molecules are encoded by multiple genes

With the exception of the β-chain for MHC class I molecules, which is encoded by the *B2M* gene on chromosome 15, each of the MHC molecule chains are encoded by multiple genes in the MHC locus. In humans, these genes are referred to as HLA genes. There are three HLA genes that encode the α-chain of MHC class I molecules (*HLA-A*, *HLA-B*, and *HLA-C*), and there are three pairs of genes that encode three MHC class II molecules (DR, DQ, and DP), with α-chains encoded by *HLA-DRA*, *HLA-DQA1*, and *HLA-DPA1*, and β-chains encoded by *HLA-DRB1*, *HLA-DQB1*,

HLA-DPB1, respectively (Figure 2.4). Some individuals carry one additional gene for the DR β-chain (one of *HLA-DRB3*, *HLA-DRB4*, or *HLA-DRB5*).

HLA genes are polymorphic

The MHC region is among the most polymorphic regions in the genome. All but one of the HLA genes (*HLA-DRA*) are highly polymorphic (see genes highlighted in yellow in Figure 2.4), with hundreds of highly divergent protein-coding sequences of each of these genes reported in the human population (Robinson et al. 2020). Variation is concentrated in regions that encode the binding groove of MHC molecules (Figure 2.5). Consequently, for each MHC molecule, an individual typically inherits a different version of the peptide binding groove from each parent, which means that most humans express six unique MHC class I molecules (corresponding to maternal and paternal versions of each of the three α-chain genes), and up to twelve MHC class II molecules (Table 2.1).

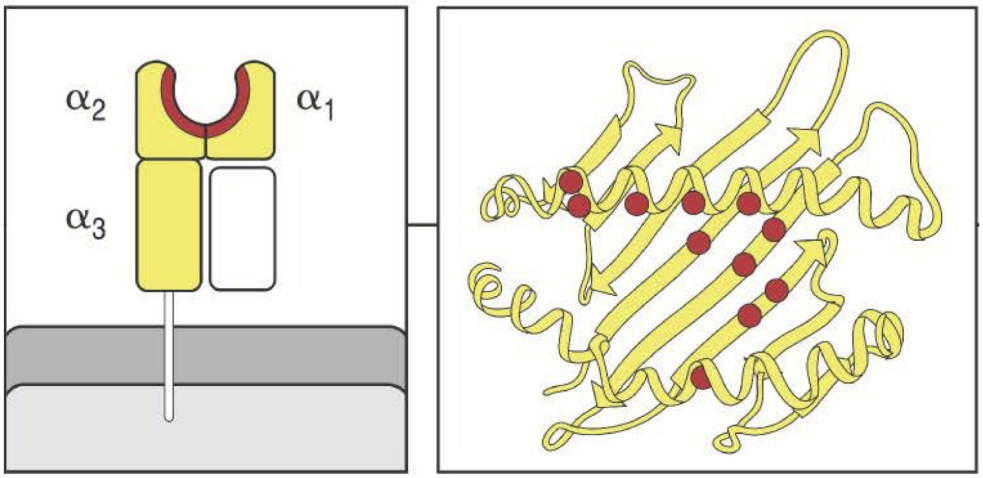

Figure 2.5: Location of genetic variation within the MHC class I molecules is indicated with red on a cartoon (left) and ribbon diagram (right) of the protein complex. Polymorphism is restricted to the α_1 and α_2 domains of MHC class I molecules. Furthermore, allelic variability within these domains is clustered in positions that line the peptide-binding groove. Figure and caption adapted from Murphy, Travers, and Walport 2012.

2.1.4 HLA typing and nomenclature

In humans, the MHC, also referred to as the HLA region, is a 4 million-basepair (Mb) region on the short arm of chromosome 6 (Figure 2.2). Over the past several decades, techniques for HLA genotyping in human subjects have evolved from antibody-based approaches (serologically-defined HLA types) to diverse molecular approaches, including restriction fragment length polymorphism (RFLP) analysis, sequence-specific oligonucleotide (SSO) hybridization, sequence-specific amplification (SSP PCR), and sequence-based typing (SBT) (Erlich 2012). Currently, the most accurate and comprehensive approach to HLA typing is sequence-based typing using either Sanger (Voorter, Palusci, and Tilanus 2014) or short-read high-throughput sequencing (Schöfl et al. 2017). However, sequence-based typing still has some challenges, for example, it can result in allelic ambiguity where the combination of reads

Table 2.1: The set of MHC class II molecules expressed within a single individual. P=paternally inherited allele; M=maternally inherited allele. Since DRA is monomorphic in human populations, we do not distinguish between paternal and maternal alleles. *Only some individuals carry a second gene encoding the DR β-chain.

α-chain allele	β-chain allele
DRA	*DRB1* (P)
DRA	*DRB1* (M)
DRA	*DRB3/4/5* (P)*
DRA	*DRB3/4/5* (M)*
DQA1 (M)	*DQB1* (M)
DQA1 (M)	*DQB1* (P)
DQA1 (P)	*DQB1* (M)
DQA1 (P)	*DQB1* (P)
DPA1 (M)	*DPB1* (M)
DPA1 (M)	*DPB1* (P)
DPA1 (P)	*DPB1* (M)
DPA1 (P)	*DPB1*(P)

observed could be derived from multiple pairs of parental haplotypes. Long-read sequencing can in theory address some of the limitations of existing sequence-based methods and may one day become the preferred approach (Mayor et al. 2015). Systems for cataloguing HLA types reflect the history of HLA typing technologies and the assumed hierarchy of genetic variation in the region, where serologically distinct subtypes and unique amino acid sequences are considered more likely to be consequential to human physiology than synonymous coding or non-coding sequence variants (Bakker and Raychaudhuri 2012).

Classical HLA types for a given HLA gene (*HLA-A*, *-B*, *-C*, *-DRB1*, *-DQA1*, *-DQB1*, *-DPA1*, or *-DPB1*) can be described at a range of resolutions: 1-field corresponds to distinct serotypes, 2-field corresponds to distinct amino acid sequences, 3-field corresponds to distinct DNA sequences in exons, and 4-field corresponds to the full DNA sequence including introns (Figure 2.6). Importantly, classical HLA types correspond to a hierarchy of gene-level haplotypes, where at the highest resolution

(4-field) each possible DNA haplotype for the entire gene corresponds to a unique allele identifier. This is in contrast to the way genetic variation is typically catalogued elsewhere in the genome, where each genetic variant is identified separately by its genomic location and nucleotide sequence relative to a reference. Due to extensive linkage disequilibrium in the MHC, the classical HLA alleles, which represent gene-level haplotypes, can be further combined into extended haplotypes spanning multiple genes, particularly for the tightly linked class II DR-DQ genes (Figure 2.7).

Figure 2.6: HLA nomenclature system. Top: Diagram showing the resolution provided by 1-, 2-, 3-, or 4-field HLA typing. Bottom: An example 4-field resolution HLA type.

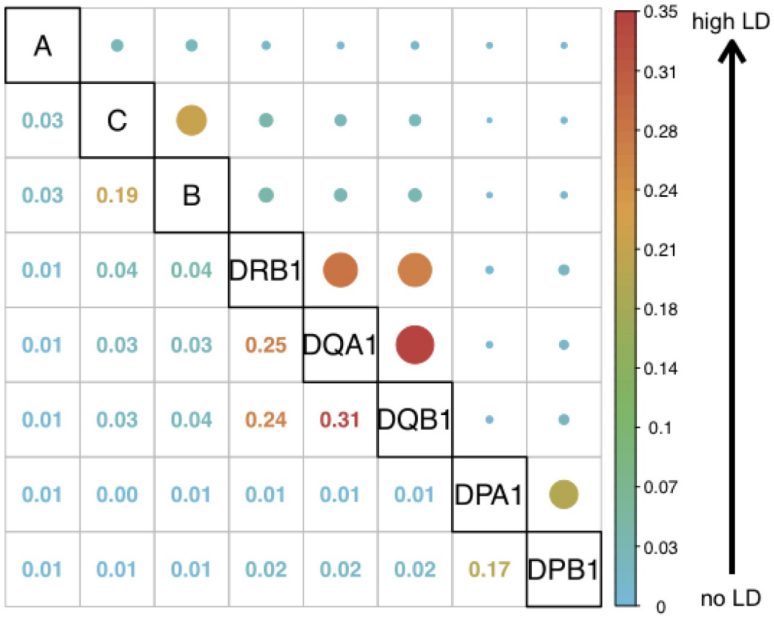

Figure 2.7: Linkage disequilibrium (LD) structure of the HLA genes. Pairwise normalized entropy (ϵ) measuring the difference of the haplotype frequency distribution for linkage disequilibrium and linkage equilibrium among five population groups. It takes values between 0 (no LD) to 1 (perfect LD). Figure and caption adapted from Luo et al. 2020.

2.1.5 Motivation

Despite substantial effort to design efficient HLA typing methods, obtaining high resolution (2-field) HLA typing for all eight HLA class I and class II genes remains laborious, sometimes involving a multi-step process combining different types of assays to resolve allelic ambiguity (Mychaleckyj et al. 2010). This is typically cost-prohibitive for large-scale genetic studies. Imputing high resolution HLA types based on SNP genotypes from the MHC region, which can be obtained at much lower costs using standard GWAS genotyping arrays, provides an opportunity to study HLA associations with common disease in large cohorts. HLA imputation methods have been developed with this aim (Karnes et al. 2017) and will be discussed in greater detail below.

While HLA associations with T1D have been well-described in European popula-

tions (Figure 2.8, also see Section 1.3.2), studies focusing on other ancestral groups have been restricted to HLA class II alleles and based on limited sample sizes (Noble et al. 2013; Howson et al. 2013). Here, we use available high resolution HLA typing from T1DGC participants (Mychaleckyj et al. 2010) of diverse ancestries to build an HLA imputation panel and impute HLA types in a larger African American T1D case-control cohort. We then analyze imputed HLA type associations with T1D to define African-ancestry HLA types contributing to T1D susceptibility.

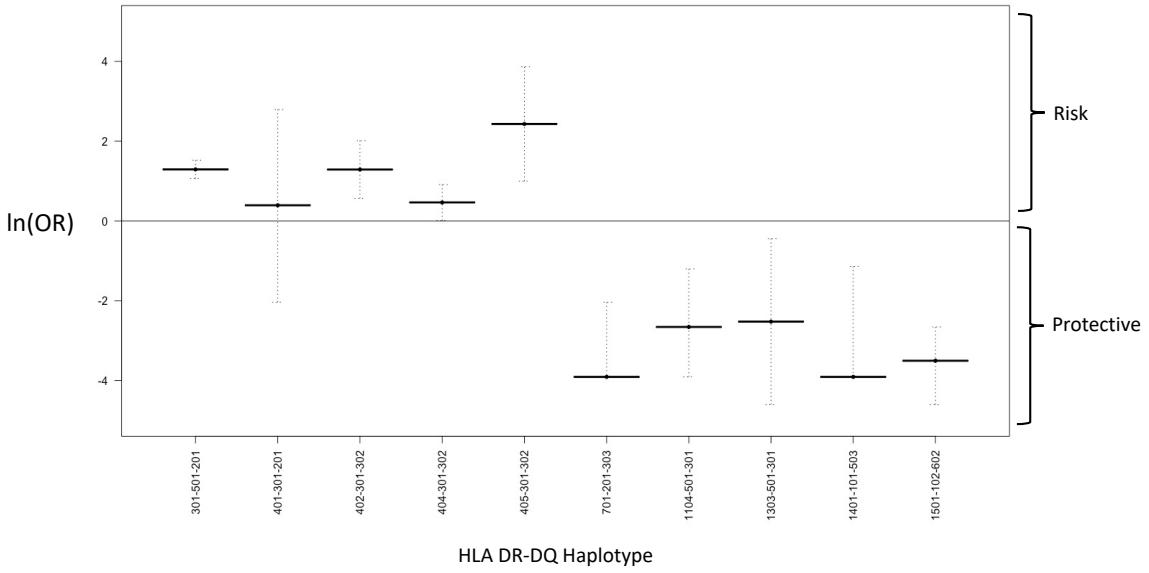

Figure 2.8: HLA DR-DQ haplotype associations with type 1 diabetes (T1D) in European ancestry families from the Type 1 Diabetes Genetics Consortium (T1DGC). Haplotypes can both increase risk and provide protection from developing T1D. Figure generated with data reported in Erlich et al. 2008.

2.2 Methods

2.2.1 Study samples and genotype generation

We obtained DNA samples and data from 666 T1D case subjects and 596 control subjects of African ancestry ascertained by the T1DGC (Rich et al. 2006), 255 case

subjects from the SEARCH for Diabetes in Youth study (SEARCH) (SEARCH Study Group 2004), 41 case subjects from the Genetics of Kidneys in Diabetes (GoKinD) study (Mueller et al. 2006), and 59 case subjects and 42 control subjects from the Barbara Davis Center (BDC) (Rewers et al. 1996). Samples were obtained from an additional 368 African-ancestry control subjects from the Consortium for the Longitudinal Evaluation of African Americans with Early Rheumatoid Arthritis (CLEAR) (Danila et al. 2017), 801 control subjects from the New York Control Population (NYCP) from the Feinstein Institute for Medical Research (Mitchell et al. 2004), 659 control subjects from the University of Alabama at Birmingham (UAB) (Li et al. 2013), and 462 control subjects from the University of California, San Francisco (UCSF) (Isobe et al. 2015). DNA from study participants was obtained after receiving approval from relevant institutional research ethics committees and informed consent.

Genotyping was performed using the ImmunoChip (Illumina), according to the manufacturer's protocols. The ImmunoChip is a custom genotyping array of about 196,000 SNPs in 186 regions associated with autoimmune diseases, including T1D. The ImmunoChip is described in greater detail in Section 3.1.1. The MHC region on the ImmunoChip consisted of about 6,000 SNPs based on an earlier focused analysis of HLA imputation in the T1DGC (Brown et al. 2009).

The T1DGC, SEARCH, GoKinD, NYCP, BDC, and UCSF samples were genotyped at the Center for Public Health Genomics, University of Virginia, Charlottesville, VA. The CLEAR and UAB control samples were genotyped at the Feinstein Institute for Medical Research, Manhasset, NY. All genotyping files were assembled at the University of Virginia to cluster genotypes using the Illumina Gentrain2 algorithm. All SNP genome positions used human reference genome GRCh37.

Sample quality control measures included call rate, heterozygosity, and concordance between reported and genotype-inferred sex. SNP quality control measures

included restricting analysis to genotyping call rates $\geq 95\%$, Hardy-Weinberg equilibrium in controls ($p > 10^{-10}$), and removal of monomorphic SNPs. To avoid cryptic relatedness that can confound association analyses, the relationship inference method implemented in KING (Manichaikul et al. 2010, www.chen.kingrelatedness.com) estimated kinship coefficients between every pair of study subjects based on ImmunoChip genotypes. In pairs of subjects found to be related, one subject was randomly removed to avoid bias.

All participants included in our study self-reported as being of African ancestry. Reference samples from the International HapMap Project (The International HapMap Consortium 2003) representing African, Asian, and European populations were used to validate self-reported ancestry via the principal component (PC) projection method implemented in KING (Manichaikul et al. 2010). Prior to PC analysis, autosomal SNPs were pruned for linkage disequilibrium ($r^2 < 0.2$) to reduce allelic correlations between adjacent SNPs. PC analysis was performed on HapMap control samples, followed by projection of our study population onto the HapMap control PC space. Study participants that self-identified as being of African ancestry and aligned with African-ancestry HapMap samples in the control space were analyzed. To control for study differences and ensure we matched case-control samples appropriately, after the initial quality control steps with HapMap samples, we removed MHC region SNPs (chr6:25,294-34,665 kb). PC analysis was performed on 2,928 control (unaffected) individuals, followed by projection of 1,021 individuals with T1D (affected) onto the control PC space. The first two PCs (PC1 and PC2) explained the majority of variance in the African-ancestry genotyping data and were included as covariates in logistic regression models.

2.2.2 Building a multi-ethnic HLA imputation panel

A subset of participants recruited through the T1DGC were HLA-typed with a PCR-based sequence-specific oligonucleotide probe system (Erlich et al. 2008; Mychaleckyj et al. 2010), providing classical 2-field HLA alleles for the eight MHC class I and II genes (*HLA-A*, *-B*, *-C*, *-DRB1*, *-DQA1*, *-DQB1*, *-DPA1*, and *-DPB1*). Many of these subjects were subsequently genotyped with the ImmunoChip, as described above. We constructed a multi-ancestry T1DGC HLA imputation reference panel based on 5,196 unrelated individuals for whom both ImmunoChip genotypes and HLA types were available, including 4,323 European-, 251 African-, 608 Asian-, and 14 "other"-ancestry individuals.

To generate the reference panel we used the software program SNP2HLA (Jia et al. 2013), which generates an imputation reference panel in the following way: First, HLA types are converted to binary markers indicating the presence or absence of each HLA allele. Next, SNP2HLA further determines the amino acid sequences corresponding to the observed HLA alleles using the EMBL-EMI Immunogenetics HLA Database (Robinson et al. 2020) and generates a set of binary markers to represent polymorphic amino acids. For multi-allelic sites, binary markers are generated for each allele. For example, if there are three possible amino acids at a position, the position would be encoded by three separate binary markers, each encoding the presence or absence of one of the possible amino acids. Finally, the genotypes across all binary markers are phased using the imputation software Beagle version 3.0.4 (Browning and Browning 2009), providing phased haplotypes for SNPs, HLA alleles, and amino acids in the HLA region (Figure 2.9).

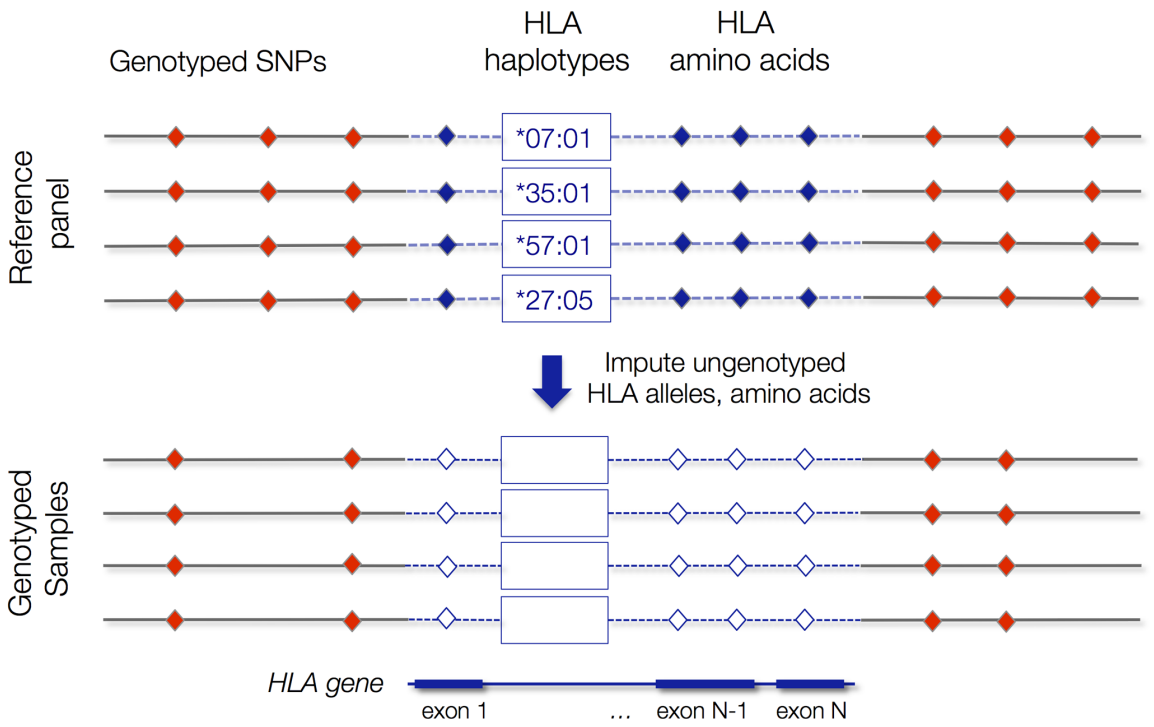

Figure 2.9: Overview of the SNP2HLA imputation procedure.The reference panel (top) contains SNPs in the MHC, classical HLA alleles at the class I and class II loci, and amino acid sequences corresponding to the 2-field HLA types at each locus. For a data set with genotyped SNPs across the MHC (bottom), we use the reference panel to impute classical alleles and their corresponding amino acid polymorphisms. Figure and caption adapted from Jia et al. 2013.

2.2.3 HLA imputation and validation

Using our multi-ethnic T1DGC HLA imputation reference panel, we imputed HLA genotypes in the remaining African-ancestry subjects. Imputation was also performed using the software SNP2HLA (Jia et al. 2013, `http://software.broadinstitute.org/mpg/snp2hla`). Imputation accuracy was empirically assessed using a validation data set of 50 randomly selected African-ancestry participants with both ImmunoChip genotypes and HLA types available. These 50 subjects were not included in the T1DGC HLA imputation reference panel because they were related to other individuals in the reference panel (the imputation reference panel intentionally included only unrelated individuals - see section 2.2.2). For the purpose of evaluating

the accuracy of imputation in this validation set of 50 individuals, any relatives of these participants were removed from the reference panel, and HLA alleles were re-imputed using the remaining reference haplotypes from 5,104 individuals (including 159 African-ancestry individuals). Imputed HLA types at each of the eight classical HLA loci in these 50 subjects were compared with their known HLA types. Imputation accuracy (ρ) for a given locus (e.g., *HLA-DRB1*) was calculated as the sum of dosages assigned to the correct alleles divided by the total number of chromosomes as previously described in Jia et al. 2013 (Equation 2.1).

$$\rho = \sum_{i=1}^{N} \frac{d_i(A_i1) + d_i(A_i2)}{2N} \tag{2.1}$$

where A_i1 and A_i2 are the true HLA types at a given locus in individual i, and $d_i(A_i1)$ and $d_i(A_i2)$ are the imputed dosages assigned to A_i1 and A_i2 in individual i, respectively. In a completely accurate imputation, $d_i(A_i1) = 1$ and $d_i(A_i2) = 1$ for every individual i.

Two additional indicators of imputed genotype quality were evaluated. First, we tested for Hardy-Weinberg equilibrium within controls in each gene, including only alleles with $MAF > 0.01$. Second, we evaluated whether the sum of dosages assigned to a given locus for each individual was close to two and excluded subjects where the sum of most likely genotypes across a locus did not equal two.

2.2.4 Modeling of HLA associations with T1D

The most likely imputed classical HLA genotypes were used in association analyses (instead of dosages or genotype posterior probabilities). We defined rare HLA alleles as those that occur fewer than 30 times in the combined sample of case and control subjects. Statistical analyses of HLA alleles were performed on 2-field imputation calls. When multiple, rare 2-field alleles (allele count < 30) were observed within the

same 1-field allelic stratum, the set of rare alleles were combined and analyzed as a single 1-field allele.

Evaluating relative contribution of variation in eight HLA genes

To assess the contribution of total variation at a given locus (e.g., *HLA-A*) to T1D risk, a multi-allelic model was fit in which all alleles for a given locus were included as independent variables. For example, for a locus with p common alleles, association between that locus and disease risk was evaluated using the model defined in Equation 2.2).

$$logit(Y_i) = \mu + \alpha_1 PC_{1i} + \alpha_2 PC_{2i} + \alpha_3 Sex_i + \beta_1 A_{1i} + \beta_2 A_{2i} + \cdots + \beta_{p-1} A_{p-1,i} \quad (2.2)$$

where Y_i is the T1D status of individual i and A_{ji} is the allele count for the jth allele at the locus in individual i. The pth allele has been arbitrarily selected as the reference allele.

Rare alleles were not included in the model and subjects carrying one or more rare alleles (1- or 2-field) at a given locus were excluded (10% of subjects contained at least one rare allele). A likelihood ratio test comparing this model to one containing only principles components and sex was used to determine statistical significance of association between the locus and disease. To test for independently associated loci, we iteratively conditioned on the most significant locus (i.e., included all alleles at that locus in the model) until no loci remained significant. Statistical significance for a locus was based on a threshold of $p < 0.00625$ (Bonferroni correction for $\alpha = 0.05$ given eight tests).

Association with individual HLA alleles

Analysis of association between T1D and individual alleles at each of the eight classical HLA genes was conducted treating each allele as a biallelic variant. The odds ratio (OR) for an allele were calculated in a logistic regression model adjusting for sex and two PCs, defined by Equation 2.3.

$$logit(Y_i) = \mu + \alpha_1 PC_{1i} + \alpha_2 PC_{2i} + \alpha_3 Sex_i + \beta_j A_{ji} \tag{2.3}$$

where A_{ji} is the allele count for the allele being tested (jth allele at the locus) in individual i and β_j is the additive effect of the allele. Thus, the odds ratio for the effect of the jth allele on T1D risk is given by $OR = \exp(\beta_j)$. Statistical significance was determined using a likelihood ratio test, comparing the HLA allele inclusive model with a reference model containing only sex and two PCs. Statistical significance for an allele was based on a Bonferroni threshold correcting for the number of HLA alleles tested with a family-wise error rate of $\alpha = 0.05$.

Association with individual HLA class II haplotypes

HLA class II haplotypes (*DRB1-DQA1-DQB1*) were inferred using phased genotypes provided by SNP2HLA (Jia et al. 2013). Each haplotype was coded as a biallelic variant, and association analyses were conducted on common haplotypes only. Odds ratios for haplotypes were calculated in a logistic regression model equivalent to that used for HLA allelic associations (Equation 2.3), except where A_{ji} is the number of copies of the jth class II haplotype being tested. Statistical significance was determined using a likelihood ratio test, comparing the HLA inclusive model with a reference model containing only sex and two PCs. Statistical significance for a haplotype was based on a Bonferroni threshold correcting for the number of class II haplotypes tested with a family-wise error rate of $\alpha = 0.05$.

Association with individual amino acid residues

Analysis of amino acid residues in HLA class I and II genes were performed using a similar approach to that used for individual HLA alleles and class II haplotypes. Specifically, each polymorphic amino acid residue was coded as a biallelic variant (presence or absence of that residue), and association analyses were conducted one residue at a time. Odds ratios for amino acid residues were calculated in a logistic regression model (Equation 2.3, where A_{ji} is the number of copies of the jth amino acid residue being tested). Statistical significance was determined using a likelihood ratio test, comparing the amino acid inclusive model with a reference model containing only sex and two PCs. Statistical significance for an amino acid residue was based on a Bonferroni threshold correcting for the number of residues tested with a family-wise error rate of $\alpha = 0.05$.

2.2.5 Conditional analyses

Multiple independent associations with HLA alleles, class II haplotypes, or amino acid residues were identified using forward stepwise regression, iteratively conditioned on the most significant association in each category until no remaining marker met the significance threshold. Specifically, conditional analysis followed the following procedure:

Round 1: Suppose there are p variants in the region. For each variant $j \in \{1 \dots p\}$, we fit a model: $logit(Y_i) = \mu + \alpha_1 PC_{1i} + \alpha_2 PC_{2i} + \alpha_3 Sex_i + \beta_j A_{ji}$. We denote the most associated variant from round 1 as $A_{(1)}$. If the association with $A_{(1)}$ meets the multiple testing corrected significance threshold for round 1, we continue to round 2.

Round 2: For each of the remaining $p - 1$ variants, we fit a new model: $logit(Y_i) = \mu + \alpha_1 PC_{1i} + \alpha_2 PC_{2i} + \alpha_3 Sex_i + \beta_{(1)} A_{(1)i} + \beta_j A_{ji}$. We denote the most associated

variant from round 2 as $A_{(2)}$. If the association with $A_{(2)}$ meets the multiple testing corrected significance threshold for round 2, we continue to round 3.

...

Round k: For each of the remaining $p-(k-1)$ variants, we fit a new model:

$$logit(Y_i) = \mu + \alpha_1 PC_{1i} + \alpha_2 PC_{2i} + \alpha_3 Sex_i + \beta_{(1)} A_{(1)i} + \cdots + \beta_{(k)} A_{(k)i} + \beta_j A_{ji}$$

We repeat this procedure until the association with the top variant, $A_{(k)}$ for round k, does not meet the multiple testing corrected significance threshold, which we define as the Bonferroni-corrected p-value threshold corresponding to a family-wise error rate $\alpha = 0.05$, correcting for the total number of markers tested at each round.

In addition to performing conditional analysis separately on sets of HLA alleles, class II haplotypes, and amino acid residues, we performed conditional analysis jointly on SNP genotypes in the MHC region, amino acid residues, and HLA alleles.

2.2.6 Design and validation of T1D genetic risk score

An African-ancestry T1D genetic risk score (GRS) was developed using ImmunoChip SNPs significantly associated with T1D ($p < 5.0 \times 10^{-8}$), including five that capture the genetic contribution of the HLA region, one from 11p15.5 (*INS* locus), and one from 17q12 (*IKZF3-ORMDL3-GSDMB* locus). The regression coefficients at each of the seven SNPs were used as individual SNP weights. A GRS for each individual was computed as the weighted sum of allele counts. The list of SNPs and weights (log(OR)) used to compute an African-ancestry GRS for T1D are provided in Table 2.2. This GRS prediction procedure was implemented in the software package KING (Manichaikul et al. 2010).

The area under the curve (AUC) from receiver operating characteristic (ROC) analysis was computed for the African-ancestry GRS by comparing the observed T1D status (case-control) with that predicted from the GRS. For internal validation, we

preformed 1,000 rounds of cross-validation: for each round, the data were randomly divided into two subsets, 80% representing a training set and the remaining 20% as the test set. For external validation, we next tested the performance of the African-ancestry GRS on an independent set of African-ancestry samples provided by Mark Atkinson, PhD, University of Florida. To assess the value of an African-ancestry GRS, we compared the performance of the African-ancestry GRS with a European-ancestry GRS when predicting T1D in the African-ancestry subjects. The European-ancestry GRS was based on our previous report (Onengut-Gumuscu et al. 2015) that forms the basis of a recently implemented T1D GRS (Oram et al. 2016). The DeLong test was used to compare the AUC difference between the two GRS models (DeLong, DeLong, and Clarke-Pearson 1988).

Performance of the GRS on prediction of T1D risk in the African-ancestry population was also compared with two polygenic risk score (PRS) models in a cross-validation procedure: a PRS model that was generated using the actual genotype data, and a PRS model that was generated using the GWAS scan summary statistics. Similar to the evaluation of the GRS, the cross-validation data sets included a training subset consisting of 80% of the samples and a test subset consisting of the remaining 20% of samples. We used the software package GCTA (Yang et al. 2011, https://cnsgenomics.com/software/gcta) to build the PRS model in the training set. For the summary-statistics-based PRS, we first ran a logistic regression adjusting for sex and two PCs of ancestry (four PCs were also investigated for sensitivity analysis), and regression coefficients at filtered SNPs (applying p-value cutoffs of 1.0, 0.5, 0.05, 5×10^{-4}, 5×10^{-6}, and 5×10^{-8} and linkage disequilibrium r^2 cutoffs of 0.2, 0.4, 0.6, and 0.8) were used to weight the genotype at the corresponding SNP. For both PRS models, a PRS was generated for each of the samples in the test set using the PRS model.

Table 2.2: T1D-associated SNPs included in the African-ancestry genetic risk score (GRS) with weights; the effect allele is the risk-increasing allele on the positive strand.

SNP	Chr	Position (bp)	Effect Allele	Odds Ratio (OR)	Weight
rs34303755	6	32450613	C	2.88	1.058
rs34850435	6	32583299	T	2.29	0.829
rs9271594	6	32591213	G	5.89	1.773
rs2187668	6	32605884	T	3.86	1.350
rs9273363	6	32626272	A	5.29	1.666
rs689	11	2182224	T	1.48	0.393
rs2290400	17	38066240	C	1.34	0.291

2.2.7 Statistical analyses

ImmunoChip SNP quality control analyses were performed using PLINK (Chang et al. 2015). Genotype-based ancestry and relationship inference were performed using KING (Manichaikul et al. 2010). The T1DGC HLA imputation reference panel was built and additional HLA types were imputed using the software SNP2HLA (Jia et al. 2013). Empirical imputation accuracy was calculated using custom Perl scripts. All statistical analyses for association in the MHC region were performed using R 3.3.1 (R Core Team 2017). Analysis code for generating HLA imputation and association results are provided at `https://github.com/ccrobertson/aa-immunochip`.

2.3 Results

After data cleaning and quality control, genotypes for 114,874 ImmunoChip SNPs in 3,949 African-ancestry participants, consisting of 1,021 participants with T1D and 2,928 control subjects, were available for analysis. This is the largest study to date of genetics of T1D in African-ancestry populations.

Table 2.3: HLA imputation accuracy at 1- and 2-field resolution.

HLA locus	1-field accuracy	2-field accuracy
A	0.964	0.916
B	0.944	0.855
C	0.998	0.946
DPA1	0.996	0.959
DPB1	0.934	0.933
DQA1	0.991	0.990
DQB1	0.996	0.973
DRB1	0.945	0.882

2.3.1 Imputation quality

Classical HLA types (1- and 2-field) and amino acid residues were imputed in all 3,949 African-ancestry participants using 5,299 ImmunoChip SNP genotypes in the MHC region and the multi-ethnic T1DGC HLA imputation reference panel as described in Section 2.2.2. Imputation yielded uniformly high imputation accuracy for 1-field classical HLA alleles ranging from $\rho = 0.934$ for *HLA-DPB1* to $\rho > 0.990$ for *HLA-DQA1*, *HLA-DQB1*, *HLA-DPA1*, and *HLA-C*) (Table 2.3). Imputation accuracy of 2-field alleles varied by locus (Table 2.3). The lowest 2-field accuracy was seen at *HLA-B* ($\rho = 0.855$) and *HLA-DRB1* ($\rho = 0.882$). Accuracy of 2-field imputation was greater than 0.91 at remaining loci, with the highest accuracy observed at *HLA-DQA1* ($\rho = 0.990$). None of the eight HLA loci deviated significantly from Hardy-Weinberg equilibrium expectations in controls. The sum of dosages across all alleles at a given locus was between 1.98 and 2.06 (expectation is 2.00) in 95% of imputations (Table 2.4). Out of 3,949 subjects, 183 were excluded from analysis because the total allele count of most likely allele calls at each locus was not equal to two. Approximately 10% of subjects contained rare alleles (1- or 2-field). Rare alleles were more common in T1D cases than controls (χ^2-test $p = 0.005$, $OR = 1.37$).

Table 2.4: Quartiles for the sum of dosages across alleles by locus.

HLA locus	$q_{0.025}$	$q_{0.5}$	$q_{0.975}$
A	2	2	2.39
B	1.89	2	2.17
C	2	2	2.18
DQA1	2	2	2
DQB1	2	2	2
DRB1	1.91	2	2.06
DPA1	2	2	2.01
DPB1	1.94	2	2.06
Combined	1.98	2	2.06

2.3.2 Allele and haplotype associations

The most significantly associated gene was *HLA-DQB1* ($p = 7.6 \times 10^{-273}$). In a forward selection-based conditional analysis, five of the eight genes associated independently with T1D. Across all 2-field HLA alleles, T1D was most significantly associated with the MHC class II allele, *HLA-DQA1**03:01 ($OR = 5.76$, $p = 7.6 \times 10^{-141}$). In general, the largest effect sizes and most significant associations were seen with DR-DQ alleles (*HLA-DRB1*, *HLA-DQA1*, and *HLA-DQB1*), which is consistent with previous studies of HLA associations with T1D in European cohort (Section Section **??**). Alleles across these three genes are highly correlated due to strong linkage disequilibrium in the region (Figure 2.7). Therefore, we evaluated association between T1D status and DR-DQ haplotypes, defined by phased alleles at *HLA-DRB1*, *HLA-DQA1*, and *HLA-DQB1*. The most significant HLA class II haplotype association with T1D was with 03:01-05:01-02:01 ($OR = 3.9$, $P = 2.6 \times 10^{-78}$). While this and other known European haplotype associations were present, several previously identified African-specific associations (Noble et al. 2013) were also confirmed (starred in Figure 2.10), including the protective African-specific DR3 haplotype 03:02-04:01-04:02 ($OR = 0.13$, $P = 4.4 \times 10^{-26}$). Univariate associations between T1D and all imputed HLA classical alleles with allele count > 30 are provided in Onengut-

Gumuscu et al. 2019 (https://doi.org/10.2337/dc18-1727, Supplementary Table 1).

In the MHC class I region, T1D was most significantly associated with *HLA-A**24:02 ($OR = 2.17$, $P = 9.8 \times 10^{-9}$), *HLA-B**15:10 ($OR = 2.21$, $P = 7.8 \times 10^{-10}$), and *HLA-C**03:04 ($OR = 1.87$, $P = 1.2 \times 10^{-10}$). The low-frequency African-specific allele *HLA-B**57:03 was protective against risk of T1D ($OR = 0.44$, $P = 1.3 \times 10^{-5}$).

In conditional analyses across all 2-field HLA class I and II alleles, eleven alleles independently associated with T1D (Table 2.5). Conditional analyses of HLA class II haplotypes identified fifteen independently associated DR-DQ haplotypes, including the African-derived risk haplotypes 09:01-03:01-02:01 ($OR = 5.75$, $P = 2.5 \times 10^{-34}$) and 07:01-03:01-02:01 ($OR = 4.69$, $P = 6.4 \times 10^{-15}$).

						Unadjusted			Erlich et al. 2008 (29)		
DRB1	*DQA1*	*DQB1*	Count	Control AF	Case AF	OR	OR	*P*	Control EU	Case EU	OR
01:01	01:01	05:01	218	0.026	0.032	1.24	1.15	0.37	0.09	0.066	0.71
01:02	01:01	05:01	288	0.040	0.026	0.65	0.66	6.7×10^{-3}	0.01	0.007	0.66
03:01	05:01	02:01	984	0.075	0.268	4.5	3.91	2.6×10^{-78}	0.125	0.341	3.64
03:02	04:01	04:02	378	0.062	0.008	0.13	0.14	4.4×10^{-26}**	—	—	—
04:01	03:01	03:01	83	0.010	0.012	1.16	0.84	0.50	0.039	0.014	0.35
04:01	03:01	03:02	215	0.012	0.072	6.4	5.36	2.4×10^{-29}	0.045	0.281	8.39
04:04	03:01	03:02	122	0.010	0.031	3.1	2.60	1.1×10^{-6}	0.032	0.05	1.59
04:05	03:01	03:02	231	0.013	0.076	6.07	6.75	1.8×10^{-40}	0.002	0.025	11.37
07:01	02:01	02:02	584	0.081	0.056	0.68	0.66	1.5×10^{-4}	—	—	—
07:01	03:01	02:01	155	0.012	0.043	3.86	4.41	7.4×10^{-18}**	—	—	—
08:04	04:01	03:01	229	0.037	0.008	0.22	0.25	7×10^{-11}**	—	—	—
08:04	05:01	03:01	61	0.010	0.002	0.2	0.23	5.9×10^{-4}	—	—	—
09:01	03:01	02:01	326	0.024	0.092	4.17	4.87	3.4×10^{-39}**	0	0.002	—
10:01	01:01	05:01	144	0.023	0.006	0.28	0.29	1.0×10^{-6}	0.007	0.003	0.49
11:01	01:02	05:02	67	0.010	0.005	0.5	0.54	0.06	—	—	—
11:01	01:02	06:02	228	0.038	0.003	0.09	0.096	5.7×10^{-19}	—	—	—
11:01	05:01	03:01	295	0.047	0.010	0.2	0.18	7.5×10^{-19}	0.065	0.012	0.18
11:02	05:01	03:01	292	0.044	0.017	0.38	0.42	2.3×10^{-7}	0.004	0.002	0.37
12:01	01:01	05:01	221	0.033	0.015	0.44	0.49	1.4×10^{-4}	—	—	—
13:01	01:03	06:03	241	0.037	0.014	0.37	0.35	7.2×10^{-9}	0.059	0.008	0.13
13:02	01:02	05:01	178	0.026	0.014	0.52	0.57	5.6×10^{-3}	—	—	—
13:02	01:02	06:04	140	0.016	0.024	1.5	1.43	0.06	0.026	0.022	0.87
13:02	01:02	06:09	224	0.031	0.023	0.75	0.79	0.17	0.003	0	0
13:03	02:01	02:01	63	0.010	0.003	0.3	0.35	4.3×10^{-3}**	—	—	—
13:03	05:01	03:01	94	0.014	0.005	0.37	0.36	4.8×10^{-4}	0.01	0.001	0.08
15:01	01:02	06:02	167	0.028	0.002	0.07	0.055	9.9×10^{-20}	0.12	0.004	0.03
15:03	01:02	06:02	699	0.115	0.015	0.11	0.13	7.1×10^{-47}**	—	—	—
16:02	01:02	05:02	82	0.011	0.009	0.8	0.93	0.80	0.001	0.001	0.74

Only haplotypes with total allele count >60, and allele frequency (AF) in case or control groups at least 1% or higher, are presented. OR estimates and *P* values generated by logistic regression, adjusting for two PCs and sex. Control AF, African-ancestry HLA haplotype frequency in subjects without type 1 diabetes. Case AF, African-ancestry HLA haplotype frequency in subjects with type 1 diabetes. Control EU, control HLA haplotype frequency from affected family-based control (AFBAC) method based upon haplotypes not transmitted to an affected child in families. Case EU, case HLA haplotype frequency from the AFBAC method based upon transmitted haplotype from a parent to a child with type 1 diabetes in families. **African ancestry–specific association with type 1 diabetes.

Figure 2.10: Association of MHC class II haplotypes with T1D in unrelated African-ancestry individuals contrasted with white families from the T1DGC

2.3.3 Amino acid associations

One of the proposed mechanisms for association between HLA polymorphism and T1D risk is altered peptide presentation to the adaptive immune system via variable amino acid sequences in the peptide binding groove of MHC class I and II molecules. Thus, analyzing the HLA region at amino acid resolution, instead of gene-level haplotypes (i.e., classical HLA alleles) or region-level haplotypes (e.g., class II haplotypes), may improve power for fine-mapping and facilitate mechanistic interpretation (Hu

Table 2.5: Eleven classical HLA alleles independently associated with T1D in African Americans. Odds ratios (OR) and p-values shown were generated by multivariable logistic regression of T1D risk, including 2 principal components, sex, and all eleven alleles as independent variables.

HLA allele	OR	p-value
*DQA1**03:01	4.83	1.3×10^{-59}
*DRB1**03:01	3.91	7.5×10^{-44}
*DQB1**06:02	0.17	1.5×10^{-23}
*DQB1**03:01	0.39	2.8×10^{-15}
*DRB1**03:02	0.22	2.4×10^{-08}
*A**24:02	2.57	2.9×10^{-08}
*DRB1**13:01	0.42	7.9×10^{-06}
*DRB1**04:05	2.08	4.5×10^{-05}
*DQB1**05:03	0.17	4.3×10^{-04}
*DRB1**10:01	0.32	8.8×10^{-04}
*A**29:02	1.89	2.4×10^{-04}

Table 2.6: Fifteen HLA class II DR-DQ haplotypes independently associated with T1D in African Americans. Odds ratios (OR) and p-values shown were generated by multivariable logistic regression of T1D risk, including 2 principal components, sex, and all fifteen haplotypes as independent variables.

DRB1	*DQA1*	*DQB1*	*OR*	p-value
03:01	05:01	02:01	3.74	4.7×10^{-49}
09:01	03:01	02:01	5.75	2.5×10^{-34}
04:05	03:01	03:02	7.66	9.6×10^{-32}
04:01	03:01	03:02	6.66	4.1×10^{-26}
07:01	03:01	02:01	4.69	6.4×10^{-15}
04:04	03:01	03:02	3.73	4.5×10^{-09}
15:03	01:02	06:02	0.21	5.6×10^{-15}
03:02	04:01	04:02	0.2	6.5×10^{-10}
15:01	01:02	06:02	0.08	1.5×10^{-06}
11:01	01:02	06:02	0.14	6.0×10^{-07}
11:01	05:01	03:01	0.25	3.1×10^{-08}
08:04	04:01	03:01	0.35	1.0×10^{-04}
10:01	01:01	05:01	0.34	8.3×10^{-04}
13:01	01:03	06:03	0.48	9.9×10^{-04}
14:01	01:01	05:03	0.17	4.8×10^{-03}

Table 2.7: Eight amino acid residues in HLA class I and II proteins independently associated with T1D in African Americans. AF= allele frequency.

HLA locus	Peptide position	Residue	Case AF	Control AF	OR	P
DQB1	57	A	0.69	0.25	5.67	1.4×10^{-157}
DQA1	47-52-54	K-H-L	0.07	0.10	0.18	3.0×10^{-45}
DRB1	11	SP	0.50	0.72	0.51	7.1×10^{-20}
DQB1	87	F	0.04	0.22	0.36	9.3×10^{-13}
DQB1	26	L	0.79	0.55	2.50	1.2×10^{-11}
A	62	QR	0.54	0.64	0.67	1.0×10^{-08}
B	158	A	0.96	0.98	0.35	3.1×10^{-07}
DQB1	30	H	0.16	0.26	1.48	2.0×10^{-04}

et al. 2015). Using SNP2HLA, we inferred amino acid sequences for each subject based on their classical HLA alleles and analyzed association between T1D and each polymorphic amino acid in the HLA class I and II genes. The amino acid association patterns across all HLA genes were largely consistent between African ancestry and Norther European ancestry subjects (Figures 2.11 and 2.12). The most significantly associated amino acid in our African American cohort was residue 57 of the DQ β-chain encoded by *HLA-DQB1* ($OR = 5.7$, $p = 2.4 \times 10^{-157}$), the same as was observed in subjects of Northern European ancestry (Hu et al. 2015). Conditional analysis revealed seven additional independent T1D associations with amino acid residues (Table 2.7, Figure 2.13), all of which have been previously identified in Northern European ancestry subjects with consistent directions of effects (Hu et al. 2015).

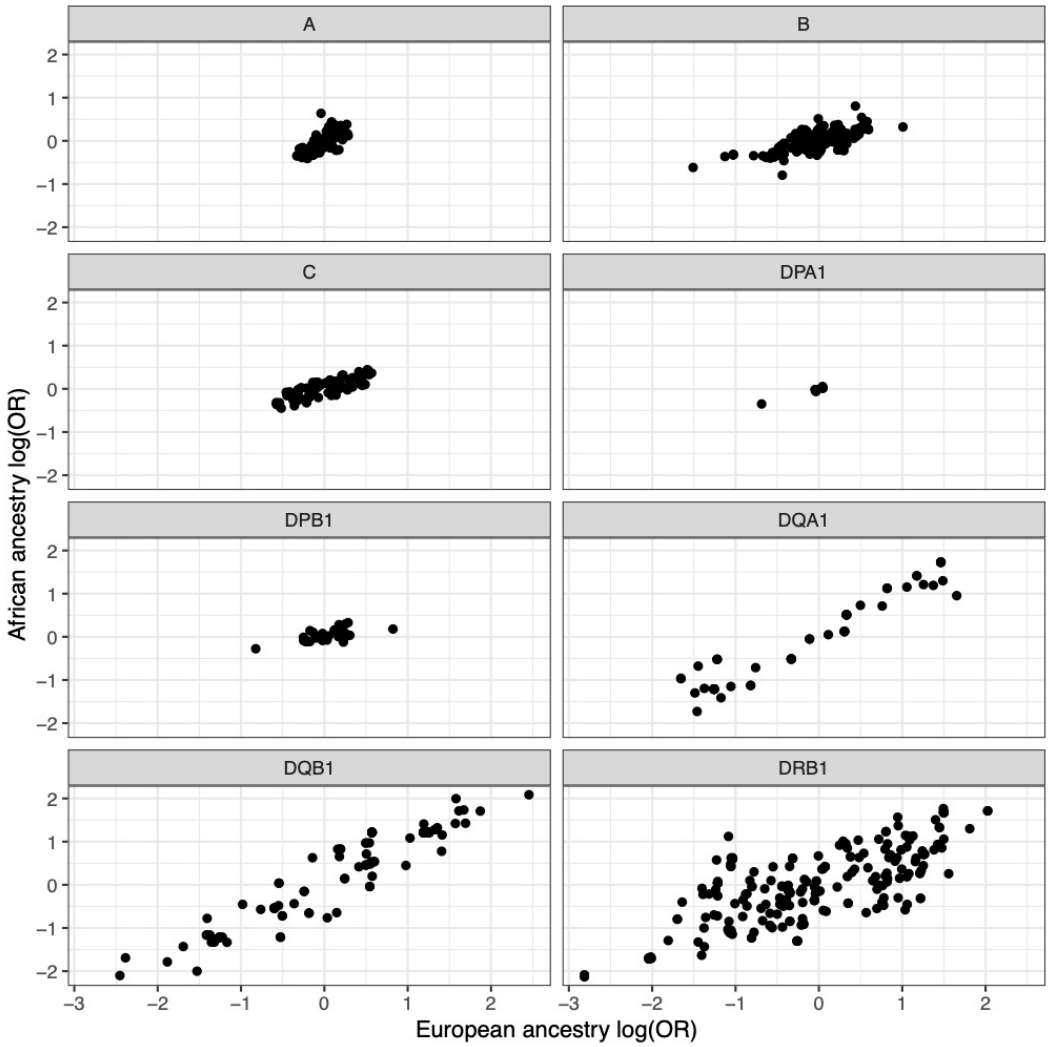

Figure 2.11: Correlation between amino acid effect on T1D risk in African- and Northern European-ancestry subjects. Northern European ancestry association statistics obtained form Hu et al. 2015.

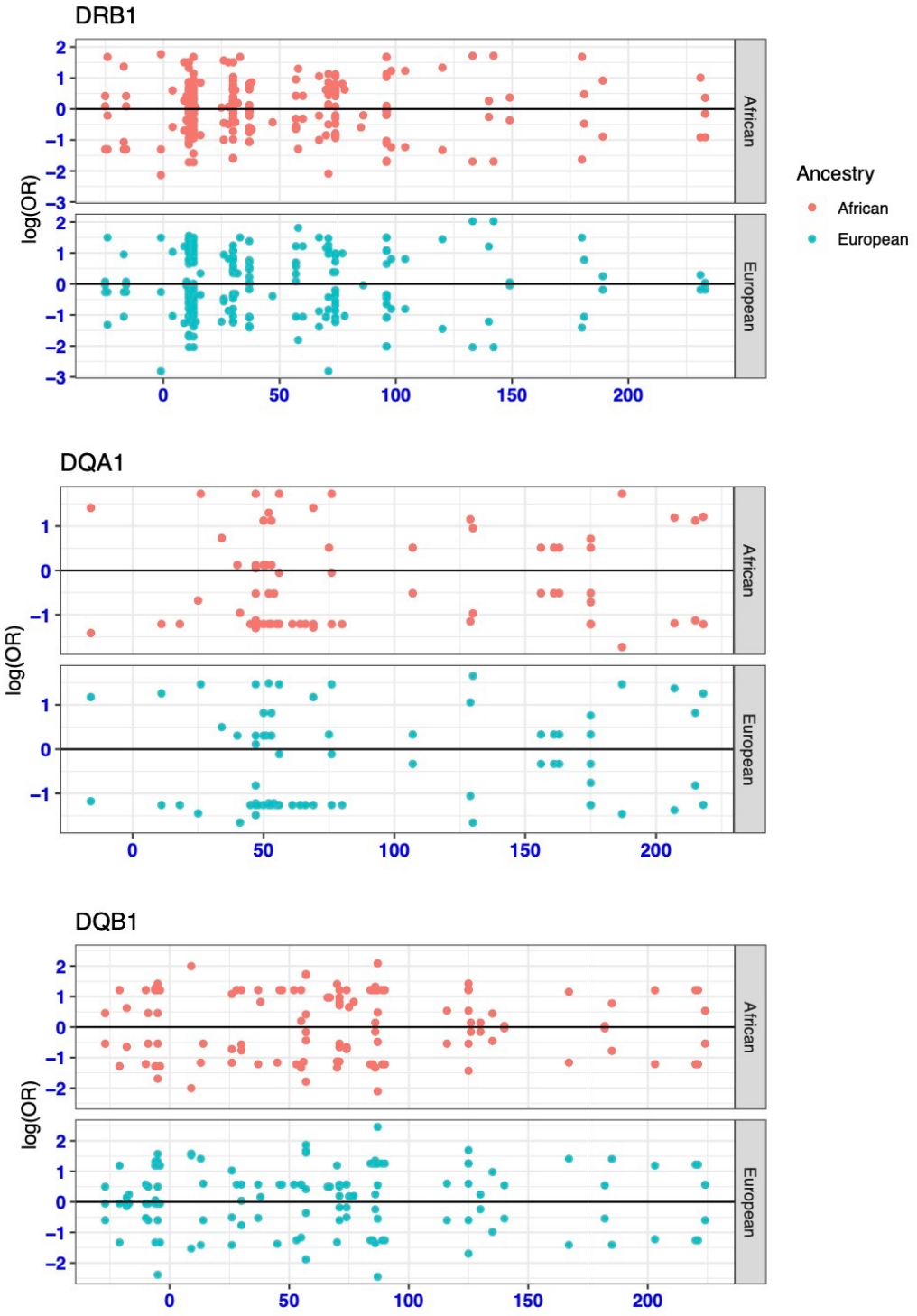

Figure 2.12: Patterns of T1D association with amino acid residues in DR-DQ proteins within African and Northern European ancestry cohorts. Northern European ancestry association statistics obtained form Hu et al. 2015.

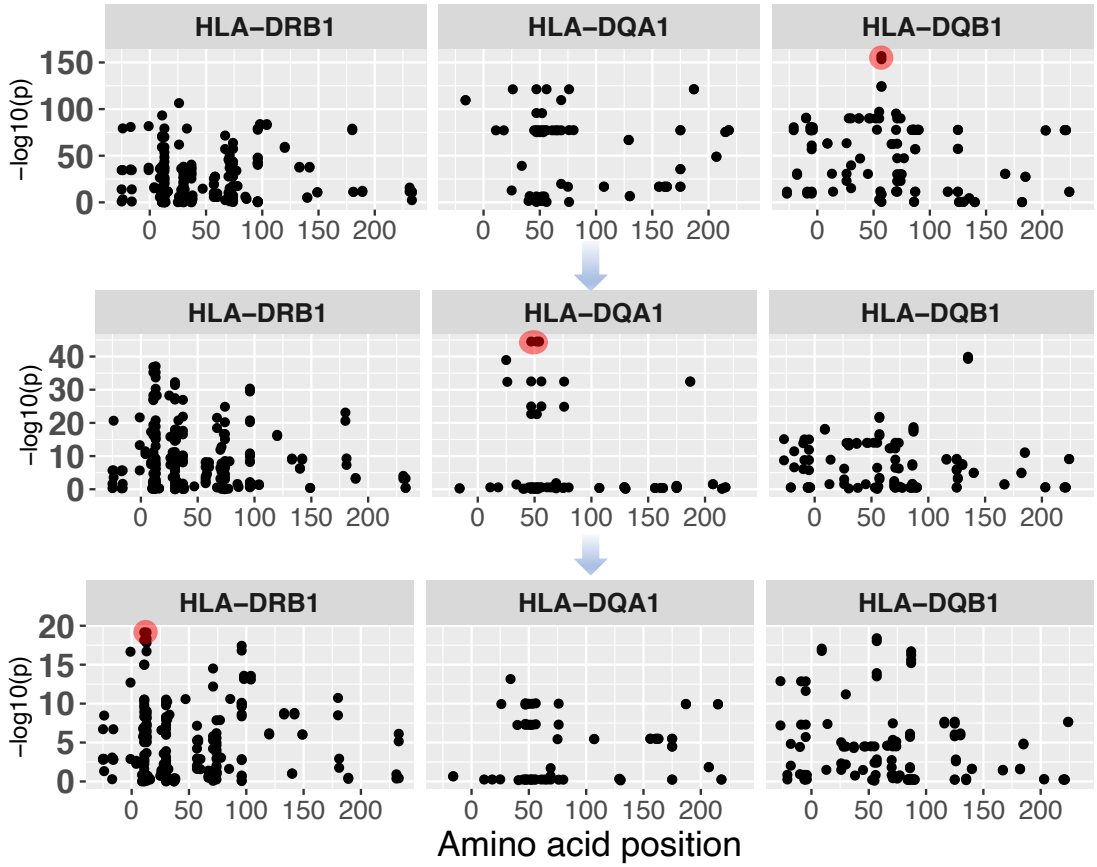

Figure 2.13: Multiple amino acid residues in class II HLA genes were independently associated with T1D in African Americans. (Top) In unconditioned association analysis, alanine at position 57 in *HLA-DQB1* is the most strongly associated residue ($OR = 5.7$, $p = 1.4 \times 10^{-157}$); (Middle) After conditioning on *HLA-DQB1* 57A, the *HLA-DQA1* 3-residue haplotype 47K-52H-54L ($OR = 0.18$, $p = 3.0 \times 10^{-45}$) was the most significant residue; (Bottom) After conditioning on *HLA-DQB1* 57A and *HLA-DQA1* 47K-52H-54L, a proline at position 11 in *HLA-DRB1* is the most associated residue ($OR = 0.51$; $p = 7.1 \times 10^{-20}$). The most significant position in each round of conditional analysis is highlighted red.

2.3.4 Improved risk prediction with African-specific HLA SNPs

In analyses led by Wei-Min Chen and Suna Onengut-Gumuscu, we assessed the performance of a previously described European-ancestry T1D genetic risk score (GRS) (Oram et al. 2016) in the African-ancestry population. The European-ancestry GRS model consisted of 30 SNPs derived from the T1DGC (Onengut-Gumuscu et al. 2015;

Noble et al. 2010) that were most associated with T1D in European populations and were polymorphic in African-ancestry populations (5 in the HLA region and 25 others). The European-ancestry GRS applied to African-ancestry samples had an AUC of 0.798, reflecting the overlap in regions of the genome associated with T1D risk. The African-ancestry T1D GRS included only seven SNPs (5 in HLA region, 1 for *INS*, and 1 in the *IKZF3-ORMDL3-GSDMB* region) yet had an AUC of 0.871 (Figure 2.14), providing a significant improvement in prediction of T1D relative to the European-ancestry GRS (($p < 2.2 \times 10^{-16}$ for comparison of AUCs). Internal validation of the African-ancestry GRS yielded similar performance in the African-ancestry samples (average $AUC = 0.870$ across 1,000 rounds of cross-validation).

To externally validate the GRS, we applied it to an independent African-ancestry cohort consisting of 61 T1D case subjects and 54 control subjects. The AUC for T1D risk prediction was 0.779. We further showed discrimination of subjects with T1D ($n = 63$) from subjects with T2D ($n = 30$) with $AUC = 0.787$. Finally, we compared the performance of our proposed African-ancestry T1D GRS and two PRSs generated in the African-ancestry population using cross-validation. The average AUC for the GRS in 100 cross-validations was 0.867, the average AUC for the genotype-based PRS was 0.808, and the average AUC for the summary-statistics-based PRS ranges from 0.797 to 0.837 (AUC is maximized at p-value cutoff 5×10^{-8} and $r^2 = 0.2$).

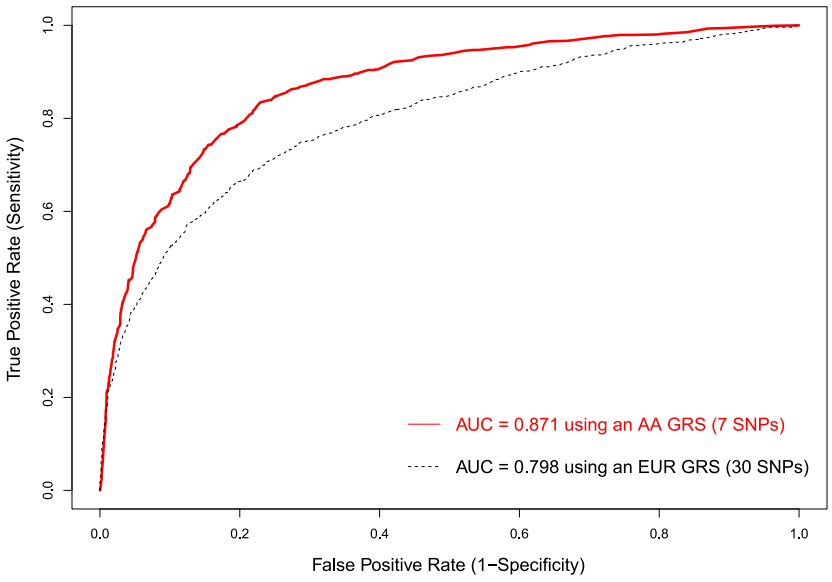

Figure 2.14: T1D risk prediction in African-ancestry subjects using a GRS. The red curve is for the prediction using an African-ancestry GRS (AA GRS), and the black curve is for the prediction using a European-ancestry GRS (EUR GRS).

2.4 Discussion

Identification of individuals at increased risk for T1D can enhance diagnostic and management practices. For example, at-risk children can be screened for presence of islet autoantibodies, which are highly predictive of disease onset. While no current intervention is available to stop progression of islet autoimmunity to clinical T1D, this information can improve surveillance, management, and education of individuals and families and provide protection from diabetic ketoacidosis (Steck et al. 2017b). This strategy has been implemented in the Fr1da study (Raab et al. 2016) and the prospective TEDDY (The Environmental Determinants of Diabetes in the Young) cohort (Bonifacio et al. 2018). Quantifying T1D risk can also help to accurately diagnose those with overlapping features of T1D, monogenic, or T2D (Oram et al. 2016). In addition to improving clinical management, effective T1D risk stratification will be critical for developing therapies to delay or prevent T1D. Identification of high-risk

individuals will enable prospective studies on early disease mechanisms and facilitate enrollment of patients into clinical trials while they are still in pre-symptomatic stages of disease. For example, while previous trials in individuals with recent onset T1D were unsuccessful, short-term treatment with an anti-CD3 monoclonal antibody delayed T1D onset by a median of 2 years among individuals with pre-symptomatic islet autoimmunity (Herold et al. 2019).

Genetic risk scores (GRS) are an emerging approach to integrating results from genetic association studies to predict human disease risk. GRS are typically created through the summation of genome-wide significant SNP genotypes, weighted by their effect sizes, into a single number that differentiates case from control status. Performance of a GRS depends on the proportion of causal variants captured by the score and the proportion of total disease risk that can be explained by genetic variation (i.e., the heritability of the disease). About 50% of T1D risk has been attributed to genetic factors. T1D is unique among complex human diseases in that the majority of genetic risk in European populations can already be explained by genome-wide significant variants. A GRS can predict T1D risk with high sensitivity and specificity in European-ancestry individuals (ROC AUC = 0.921 in Sharp et al. 2019). For this reason, one can already envision using a T1D GRS in study design and clinical practice.

Although the highest incidence of T1D has historically been in European-ancestry populations, in the U.S., incidence rates are increasing rapidly in individuals with African and Hispanic ancestry (Mayer-Davis et al. 2017). Thus, if the emerging paradigm for clinical management, clinical trial design, and prospective research studies is to rely on accurate characterization of genetic risk for T1D, we urgently need to develop genetic screening tools that will be effective across all ancestral populations. To achieve this, we must thoroughly define the genetic contributions to disease risk in historically understudied populations, including African and Hispanic

ancestry groups. Despite extensive genetic research in T1D, data from populations of non-European ancestry remain limited. In this report, we have assembled and genotyped the largest collection of African-ancestry T1D cases to date. Using this unique resource, we explore the genetic basis of T1D risk in African Americans, with a particular focus on defining African-specific HLA alleles contributing to T1D risk.

Our analyses support the strong impact of HLA class II alleles and haplotypes on T1D risk in African-ancestry groups, as previously reported (Noble et al. 2013; Howson et al. 2013). In addition, we identified new significant associations in HLA class I genes (*HLA-A*, *HLA-B*, and *HLA-C*) with T1D in African-ancestry populations. The majority of HLA associations with T1D in African-ancestry subjects are concordant with those observed in European-ancestry populations (Erlich et al. 2008). Nonetheless, we found that an African-ancestry T1D GRS performs significantly better than a previously reported European-ancestry T1D GRS in our African-ancestry cohort. Together, these findings support a model where genetic mechanisms of disease are largely consistent across populations but population-specific genetic architecture, influenced by allele frequency and linkage disequilibrium patterns, may affect the performance of genetic risk models. Thus, population-specific or multi-ancestral T1D GRS models can provide significantly improved prediction and opportunities for targeted interventions in T1D.

Chapter 3

Discovery and fine mapping of type 1 diabetes loci using the ImmunoChip

This chapter is adapted from:

> **Robertson CC**, Inshaw JRJ, Onengut-Gumuscu S, Chen WM, Flores Santa Cruz D, Yang H, Cutler AJ, Crouch DJM, Farber E, Bridges SL Jr., Edberg JC, Kimberly RP, Buckner JH, Deloukas P, Divers J, Dabelea D, Lawrence JM, Marcovina S, Shah AS, Greenbaum CJ, Atkinson MA, Gregersen PK, Oksenberg JR, Pociot F, Rewers MJ, Steck AK, Dunger DB; Type 1 Diabetes Genetics Consortium; Wicker LS, Concannon P, Todd JA, and Rich SS. Fine-mapping, trans-ancestral and genomic analyses identify causal variants, cells, genes and drug targets for type 1 diabetes. *Nature Genetics*. 2021 Jun 14:1-0.

Supplementary tables referenced in this chapter can be obtained at: https://doi.org/10.1038/s41588-021-00880-5.

3.1 Background

3.1.1 The ImmunoChip

The ImmunoChip is an Illumina Infinium custom genotyping array designed to facilitate discovery and fine mapping by providing dense genotyping of genetic regions previously implicated in immune-mediated diseases (with either suggestive or statistically significant evidence) (Cortes et al. 2011; Jostins 2013). The array was designed by a consortium of investigators to cover regions of interest for rheumatoid arthritis, ankylosing spondylitis, systemic lupus erythematosus, T1D, autoimmune thyroid disease, celiac disease, multiple sclerosis, ulcerative colitis, Crohn's disease, and psoriasis (Jostins 2013). The consortium identified 290 disease-associated regions from a collection of previous publications and pre-publication analyses, which were distilled into 188 non-overlapping "ImmunoChip regions" (Figure 3.1, Supplementary Table 2) for the analyses presented in this chapter. Within these immune-related regions, all known SNPs (as of the 1000 Genomes Project data release in February 2010) were included on the array. In addition, for each of eleven non-immunological diseases studied in the Wellcome Trust Case-Control Consortium 2 (*WTCCC2 Studies*), about 3,000 SNPs were included at previously identified candidate genes. Finally, the array included dense genotyping of SNPs in the MHC region on chromosome 6, including about 6,000 SNPs based on an earlier analysis (Brown et al. 2009), and the region encoding KIR alleles on chromosome 19. In total, the array included 196,524 variants.

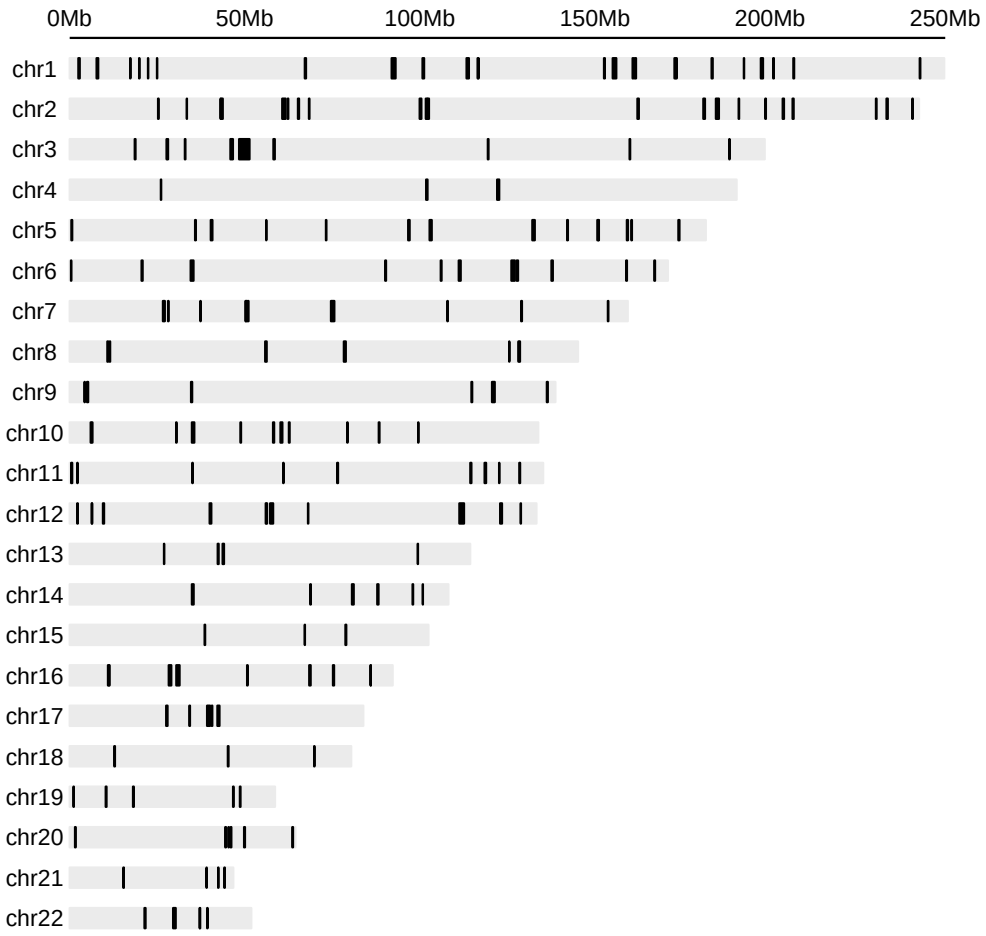

Figure 3.1: Visual representation of the 188 "ImmunoChip regions."

3.1.2 Genotype imputation

DNA genotyping arrays are a cost-effective way to genotype hundreds of thousands of genetic variants simultaneously. However, DNA arrays still only capture a small fraction of the genetic diversity in human populations. Even the largest DNA arrays, which can contain up to a few million genetic variants, capture less than 20% of common genetic variation.

Linkage disequilibrium occurs when alleles at distinct loci co-segregate during meiotic recombination, and therefore are associated in a population (Slatkin 2008).

Due to linkage disequilibrium, genotypes at variants located physically near to one another (< 50kb) are correlated (Figure 3.2). This feature of human genomes can be used to infer genotypes at genetic variants that were not included on a genotyping array using a reference panel of phased haplotypes (Figure 3.3). Imputation of genetic variants using large reference panels, including the 1000 Genomes Phase 3 (1000 Genomes Project Consortium 2015) and TOPMed (Taliun et al. 2021) haplotype panels, can improve power for genetic discovery and fine mapping.

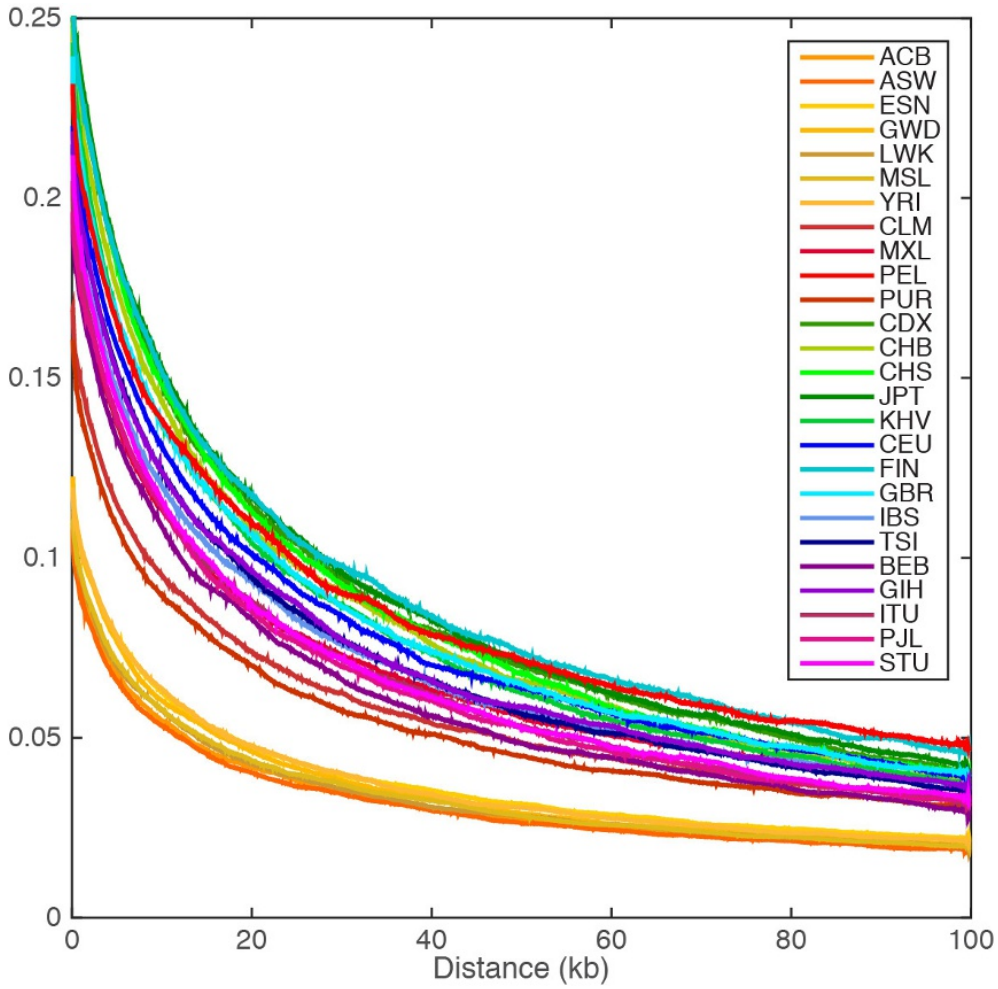

Figure 3.2: Decay of linkage disequilibrium as a function of physical distance. Linkage disequilibrium was calculated around 10,000 randomly selected polymorphic sites in each population, having first thinned each population down to the same sample size (61 individuals). The plotted line represents a 5 kb moving average. Figure and caption from 1000 Genomes Project Consortium 2015.

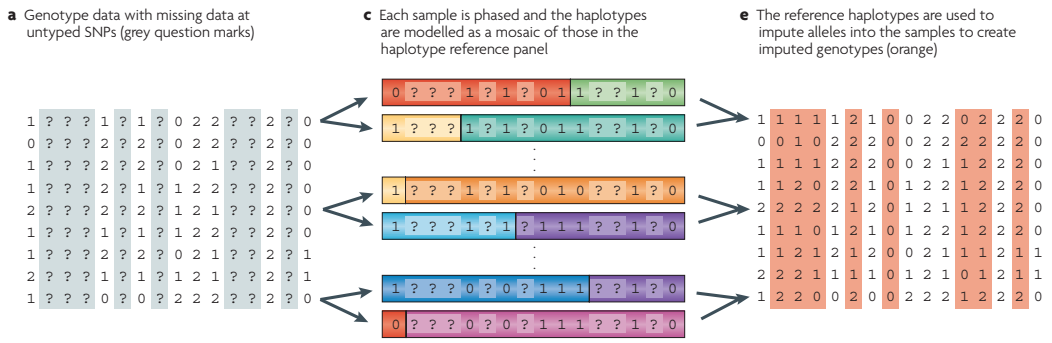

Figure 3.3: Diagram demonstrating genotype imputation using a reference panel. Figure adapted from Marchini and Howie 2010.

A number of statistical methods for genotype imputation using haplotype reference panels have been developed and improved over time. Current versions of the most popular tools include IMPUTE 5 (Rubinacci, Delaneau, and Marchini 2020, https://mathgen.stats.ox.ac.uk/impute/impute.html), Beagle 5.1 (Browning, Zhou, and Browning 2018, https://faculty.washington.edu/browning/beagle/b5_1.html), and Minimac4 (Das et al. 2016, https://genome.sph.umich.edu/wiki/Minimac4). These approaches use Hidden Markov Models to model haplotypes from genotyped individuals as a mosaic of reference haplotypes (Marchini and Howie 2010; Li and Stephens 2003). To increase the computational efficiency and facilitate imputation with extremely large haplotype reference panels, current approaches to imputation "pre-phase" the genotyped samples (Howie et al. 2012). More recently, due to the substantial computational burden and bioinformatic skills required to implement large-scale imputation, as well as privacy concerns associated with broad distribution of haplotype reference panels (e.g., the TOPMed reference panel includes haplotypes from nearly 100,000 individuals), the human genetics community has developed imputation servers (https://imputation.biodatacatalyst.nhlbi.nih.gov, https://www.sanger.ac.uk/tool/sanger-imputation-service), which allow researchers to impute genotyped samples for free and without advanced computing

resources.

For each genotyped sample, the imputation method returns a posterior probability for each possible genotype at every polymorphic position in the reference panel. These posterior probabilities can then be used in association analyses, with tools designed to accommodate these inputs (Marchini and Howie 2010), or converted to imputation "dosages" (Equation 3.1), which can then be treated like a typical independent variable in linear models.

$$D_{ij} = 0 \times Pr(G_{ij} = \text{A/A}) + 1 \times Pr(G_{ij} = \text{A/B}) + 2 \times Pr(G_{ij} = \text{B/B}) \tag{3.1}$$

where D_ij is the dosage for allele B at variant i in individual j and $Pr(G_{ij} = \text{A/A})$, $Pr(G_{ij} = A/B)$, and $Pr(G_{ij} = B/B)$ are the imputation posterior probabilities that individual j carries 0, 1, or 2 copies of the B allele at variant i, respectively.

In addition to returning imputed genotype probabilities, imputation methods also return metrics of genotype imputation quality for each variant. Minimac4, which was used for analyses in this chapter, returns an estimated "imputation R-squared" statistic, which is calculated as (https://genome.sph.umich.edu/wiki/Minimac_Diagnostics):

$$\hat{r}^2 = \frac{Var(D_{i*})}{\hat{p}(1-\hat{p})} \tag{3.2}$$

where $Var(D_{i*})$ is the variance of the vector of dosages at variant i across all individuals, and p is the allele frequency of variant i. Only imputed variants with sufficient imputation quality, as defined by some filtering threshold (e.g., imputation R-squared > 0.3), are included in downstream analyses.

3.1.3 Statistical methods for fine mapping

Statistical fine mapping of genetic association loci is one approach to differentiate potentially causal variants underlying an association from benign variants showing association merely due to linkage disequilibrium with causal variants. This challenge can be framed, statistically, as a variable selection problem (Figure 3.4), with several important features. First, the input is high dimensional, with typically hundreds to thousands of genetic variants to consider in a region. Second, it is assumed that only one or a handful of variants in a region are causally associated with disease, thus the solution is typically sparse. Third, due to linkage disequilibrium, independent variables will be highly correlated, often including variants in perfect correlation ($r^2 = 1$) with one another. In many statistical models, this will cause model instability. Finally, we are interested in inference, not prediction. In a prediction setting, one can randomly select between tightly or perfectly correlated independent variables without affecting predictive performance. In fine mapping, it is important for the solution to reflect the fact that any one of a group of tightly correlated variables have a similar likelihood of being causally related to the disease.

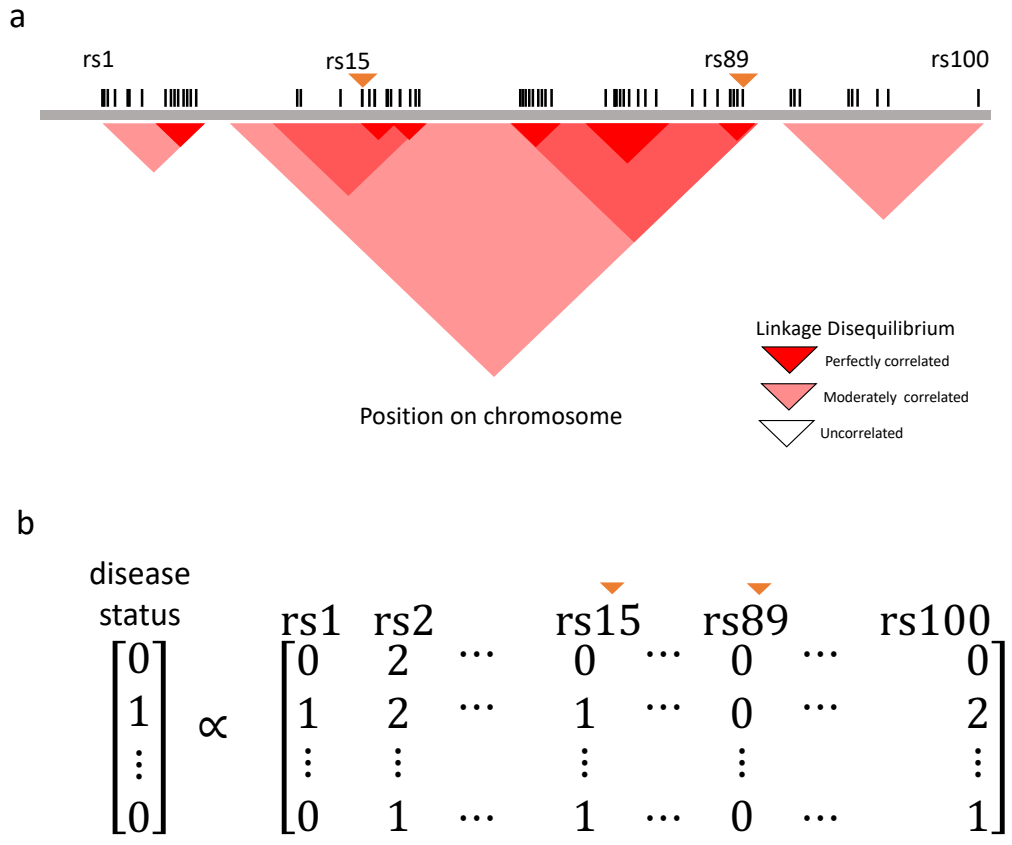

Figure 3.4: Framing fine mapping as a variable selection problem: (a) Hypothetical region of the human genome containing 100 common genetic variants, labeled rs1 through rs100; rs15 and rs89 are causal variants contributing to disease risk; however, due to extensive linkage disequilibrium in the region, many of the other variants in the region are also associated with the disease. (b) Matrix representation of the statistical fine mapping problem for the region in (a); determining the most likely causal genetic variants is a problem of assigning a probability of causality to each independent variable in the matrix.

Many statistical methods have been developed to address this problem in human genetics. The simplest, and most widely-used, method for determining the number of independent causal variants in a region is forward stepwise regression ("conditional analysis"). From independent variants identified by this method ("index variants"), credible sets can be constructed (Maller et al. 2012) by calculating the posterior probability (PP) for variant j as:

$$PP_j = \frac{BF_j}{\sum_k (BF_k)} \tag{3.3}$$

where BF_j is the Bayes factor for variant j and $\sum_k (BF_k)$ is the sum of Bayes factors for all variants in the region. Specifically, BF_j can be thought of as

$$BF_j = \frac{Pr(data|H_1)}{Pr(data|H_0)} \tag{3.4}$$

where H_1, the alternative hypothesis, is represented by an association model containing variant j, and H_0, the null hypothesis, is represented by an association model without variant j. In regions with multiple independent index variants, credible sets can be constructed for each index variant by conditioning on all other index variants during model fitting for both H_0 and H_1. Then, the 95% credible set for an index variant is the smallest set of variants for which the total posterior probability is at least 0.95. While this approach is conceptually straightforward and easy to implement, it can fail to identify causal variants in regions with complex genetic architecture. In the example shown in Figure 3.4, with multiple causal variants, this forward stepwise selection approach may prioritize variants correlated with both of the causal variants, but not the causal variants themselves. Additional modeling approaches have been developed to account for such complex situations (Table 3.1), most using Bayesian approaches to assign posterior probabilities to all variants in an associated region.

Table 3.1: Statistical methods for genetic fine mapping.

Method	Citation	Algorithm	Model input
CAVIAR	Hormozdiari et al. 2014	Exhaustively explores models with up to 6 causal variants	Genetic association summary statistics and a reference linkage disequilibrium matrix
FINEMAP	Benner et al. 2016	Approximate exhaustive approach, targetting the best combinations of variants	Genetic association summary statistics and a reference linkage disequilibrium matrix
GUESSFM	Wallace et al. 2015	Evolutionary stochastic search as implemented in GUESS (Bottolo and Richardson 2010)	Full data set including individual-level genotypes
SuSiE	Wang et al. 2020	Iterative Bayesian stepwise selection	Either the full data set or summary statistics

In this chapter, a method called GUESSFM (Wallace et al. 2015) is used, which implements the following procedure (Figure 3.5):

> The $(n \times m)$ genetic matrix X at a locus, where n is the number of individuals and m is the number of variants, is pruned to remove variants in high linkage disequilibrium($r^2 \geq 0.99$), generating a pruned $(n \times p)$ matrix, Z, with p "tag" variants. Now the model space contains 2^p possible models. A stochastic search, implemented by GUESS (Bottolo and Richardson 2010), is carried out across these 2^p models, averaging over the other parameters (e.g., variant effect size), to obtain a selection of models with high marginal posterior probabilities. This set of models is

then expanded to include models where the tag variant is replaced by each of the variants in high linkage disequilibriumwith it (which had been removed during pruning prior to the stochastic search). For each model, an Approximate Bayes Factor (ABF) is calculated by treating the binary outcome (T1D status) as linear and using the linear regression Bayesian Information Criterion (BIC). The ABF for each model can be interpreted as the support for the model relative to a null model with no genetic variants. To obtain the posterior probability for each model, the ABF is multiplied by the prior, which we took in all cases to be a binomial prior with $3/m$ expected variants included in the model, divided by the normalizing factor, the sum of all tested model posterior probabilities. The marginal probability for each SNP is taken as the sum of the posterior probabilities for all models in which it is present.

The success of any fine mapping method will depend on the linkage disequilibrium structure of the region to which it is applied, and in particular, how many variants are in very tight linkage disequilibrium with the true causal variants. For example, if a causal variant is in perfect linkage disequilibrium with multiple nearby variants (those variants are always inherited together), then no statistical fine mapping method will be able to prioritize one variant over another based on genotyping data alone. Incorporating multiple ancestry groups into genetic studies, which often have differing patterns of linkage disequilibrium (Figure 3.6), can in theory improve fine-mapping resolution (Wojcik et al. 2019a). However, the appropriateness of this approach depends on the local genetic architecture within each population. For example, defining a credible set by integrating summary statistics from two populations assumes the same variant(s) are causal in both populations. Thus, it is important to have sufficient sample sizes and statistical power to effectively evaluate this assumption in each ancestry group.

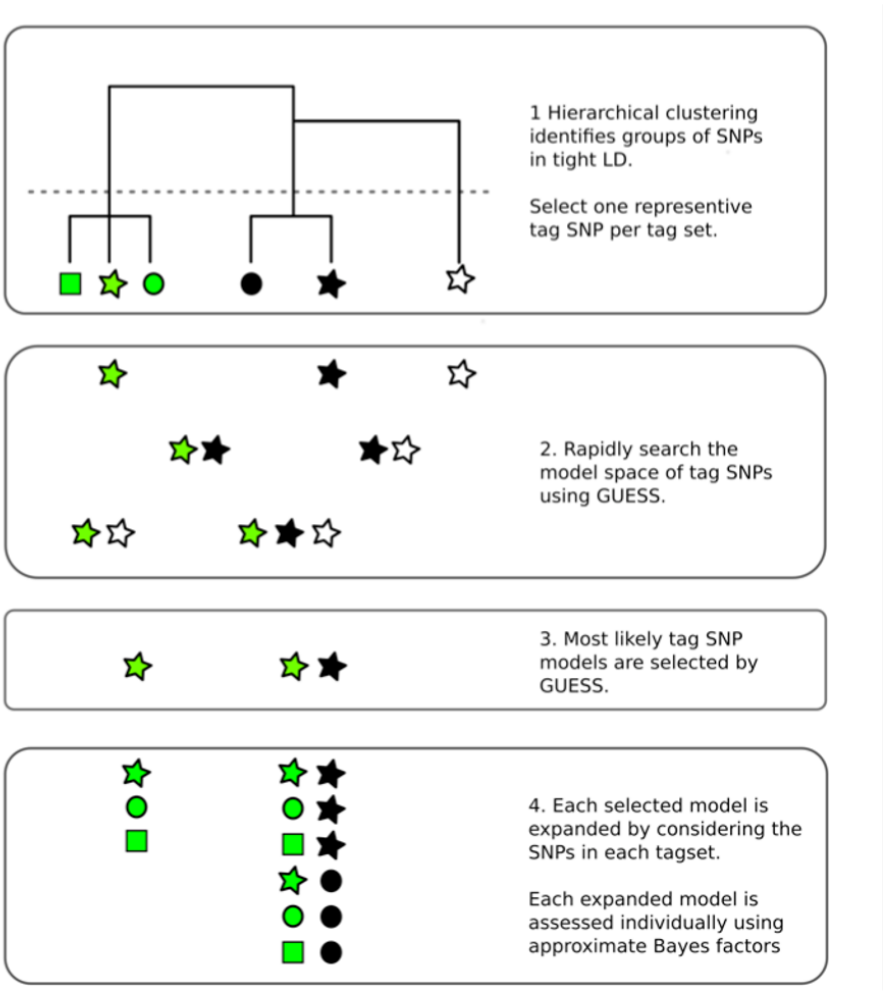

Figure 3.5: GUESSFM procedure. Figure adapted from Wallace et al. 2015.

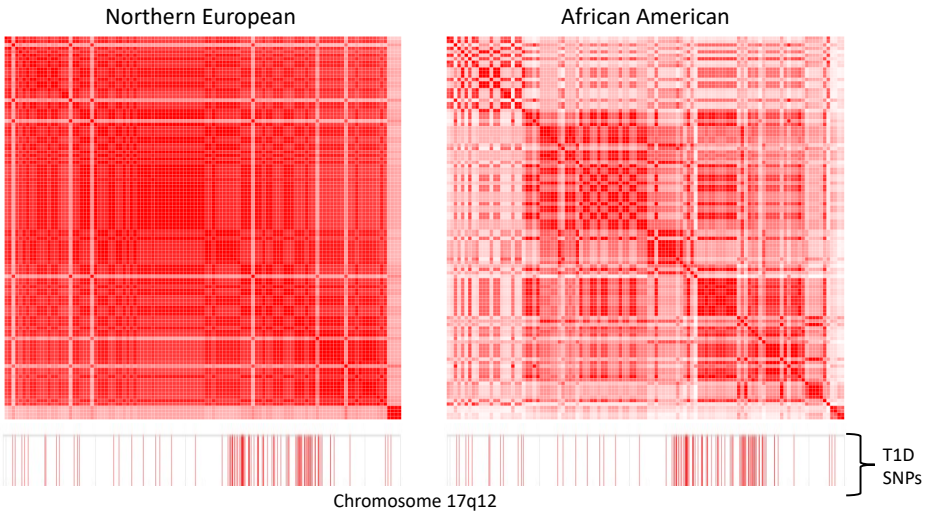

Figure 3.6: A region on chromosome 17 where T1D-associated variants have dramatically different linkage disequilibrium patterns between African- and European-ancestry 1000 Genomes populations. Figure generated at https://ldlink.nci.nih.gov.

3.1.4 Motivation

Approximately 60 genomic regions have been associated with T1D risk in individuals of European ancestry (Todd et al. 2007; Wellcome Trust Case Control Consortium 2007; Cooper et al. 2008; Hakonarson et al. 2008; Grant et al. 2009; Barrett et al. 2009; Bradfield et al. 2011; Huang et al. 2012; Onengut-Gumuscu et al. 2015; Zhu et al. 2019). However, less is known in non-European ancestry groups, despite recent increases in T1D diagnoses in these understudied populations (Section 1.1.2). Additionally, due to linkage disequilibrium, causal variants are unknown at most T1D-associated loci. Here, we double the sample size from the previous largest T1D study (Onengut-Gumuscu et al. 2015), genotype ancestrally diverse T1D cases, controls, and affected families, and impute additional variants. Using this expanded data set, we perform discovery and fine-mapping analyses.

3.2 Methods

3.2.1 Genotyping and quality control

The DNA samples were genotyped on the Illumina ImmunoChip at the University of Virginia's Genome Sciences Laboratory in the Center for Public Health Genomics ($n = 52,219$), Sanger Institute ($n = 4,347$), University of Cambridge ($n = 2,941$) and Feinstein Institute ($n = 1,811$). Raw genotyping files were assembled at UVA. Genotype clusters were generated using the Illumina GeneTrain2 algorithm. Stringent SNP- and sample-level quality-control filtering and data cleaning were performed to ensure high-quality genotypes and accurate pedigrees (Figure 3.7). In addition, the following variant filtering steps were performed:

1. Re-annotated ImmunoChip variant positions by aligning probe sequences to GRCh37 and the removing any variants with $< 100\%$ match or multiple matches

at different positions in the genome;

2. Removed variants with call rates $< 98\%$;

3. Removed variants with any discordance between duplicate or monozygotic twin samples, as confirmed by genotype-inferred relationships;

4. Removed variants with Mendelian inconsistencies in $> 1\%$ of the informative trios or parent-offspring pairs, based on genotype-inferred relationships.

For sample filtering, X-chromosome heterozygosity and Y-chromosome missingness were used to identify and exclude participants with apparent sex-chromosome anomalies or resolve inconsistencies with the reported sex. Pedigree-defined and genotype-inferred sample relationships were compared using KING version 2.1.3 (Manichaikul et al. 2010, www.chen.kingrelatedness.com). Samples were excluded when inconsistencies could not be resolved, including relationships between families, within and across cohorts. For each pair of related families observed, one family was randomly excluded from the association analysis. After resolving sex and relationship issues, samples with a genotype call rate $< 98\%$ were removed. Variants with genotype frequencies deviating from Hardy-Weinberg equilibrium ($p < 5 \times 10^{-5}$) in unrelated European-ancestry controls were excluded before imputation.

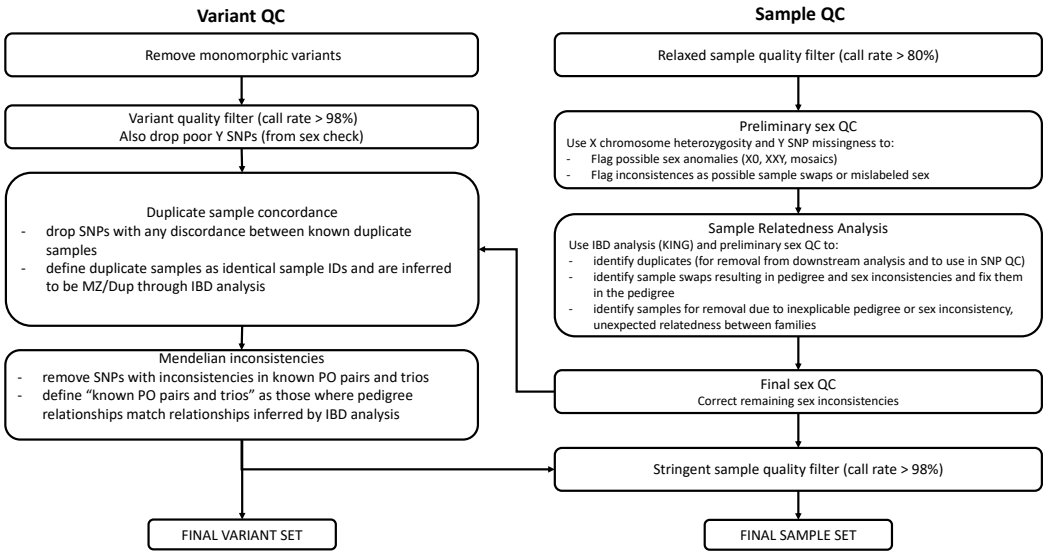

Figure 3.7: Sample and variant quality control pipeline prior to imputation.

3.2.2 Stratification of major ancestry groups and family trios

Principal components were generated in 1000 Genomes phase 3 individuals using 8,297 autosomal ImmunoChip variants selected by excluding regions of long-range linkage disequilibrium(Price et al. 2008), pruning for short-range linkage disequilibrium($r^2 < 0.2$ in 50-kb windows) and filtering for MAF > 0.05. The participant genotypes were projected onto the 1000 Genomes principal-component space using PLINK v1.9 (Chang et al. 2015, www.cog-genomics.org/plink). The first ten principal components were used in k-means clustering to define clusters of ancestrally similar participants, which were labeled according to their closest 1000 Genomes super-population. Two distinct clusters mapped to the European super-population, one of which almost exclusively consisted of participants recruited through Finnish sites. Although visualization with only the first two principal components showed overlap between Finns and other European groups, Finnish individuals consistently clustered away from the remaining European individuals when clustering used additional principal components, likely due to the unique population history of Finland (Jakkula et al.

2008). These results suggested sufficient substructure to warrant additional stratification into Finnish and non-Finnish Europeans. Thus, for association analysis, participants in this study were stratified into five ancestrally-similar groups: African admixed ("AFR"), East Asian ("EAS"), Finnish ("FIN"), other European ("EUR"), and other Admixed ("AMR") (Figure 3.8).

For case-control analyses to be performed within each ancestry cluster, affected trios were excluded and a set of unrelated individuals was selected from the remaining subjects using the "unrelated" command in KING version 2.1.3 (Manichaikul et al. 2010). Cluster-specific principal components were calculated by performing principal-component analysis on unrelated controls and projecting the remaining subjects onto the resulting axes. The remaining population stratification within each ancestry cluster was assessed visually (Figure 3.9).

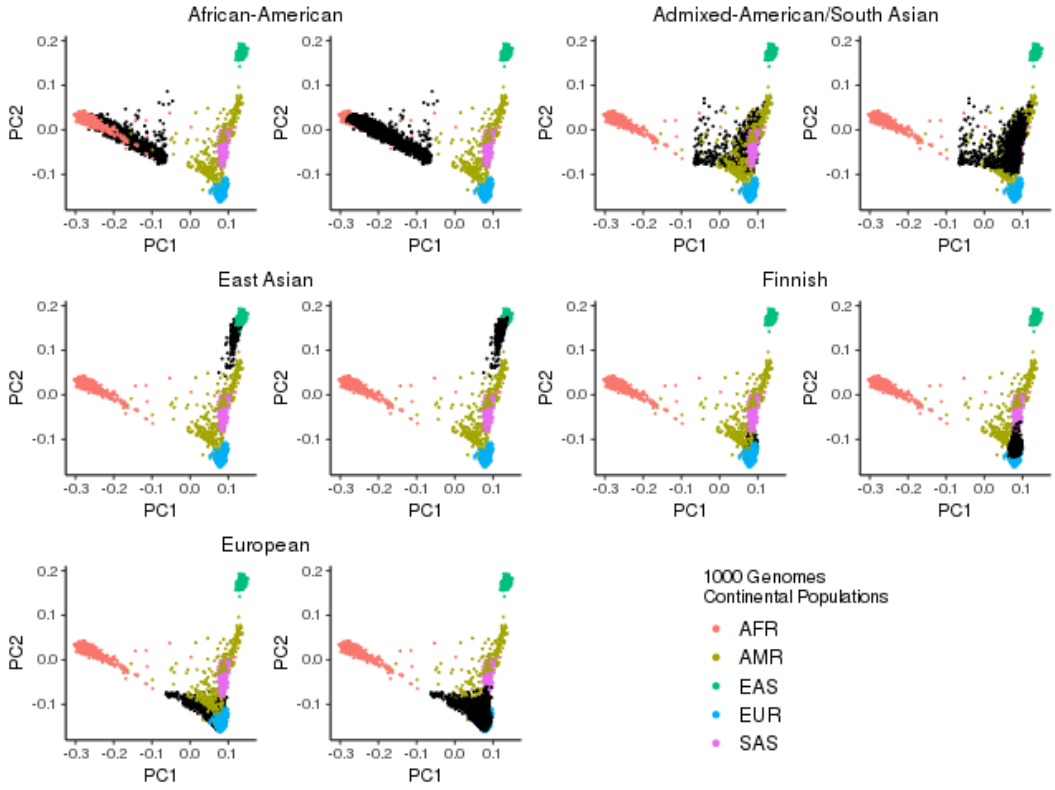

Figure 3.8: Population structure in the analysis cohort. Two plots are shown for each stratification group, one with the 1000 Genomes Project data on top (left), the other with the study participants on top (right). Study participants, black; 1000 Genomes participants, colored by super-population. AFR=African, AMR=Admixed American, EAS=East Asian, EUR=European, SAS=South Asian.

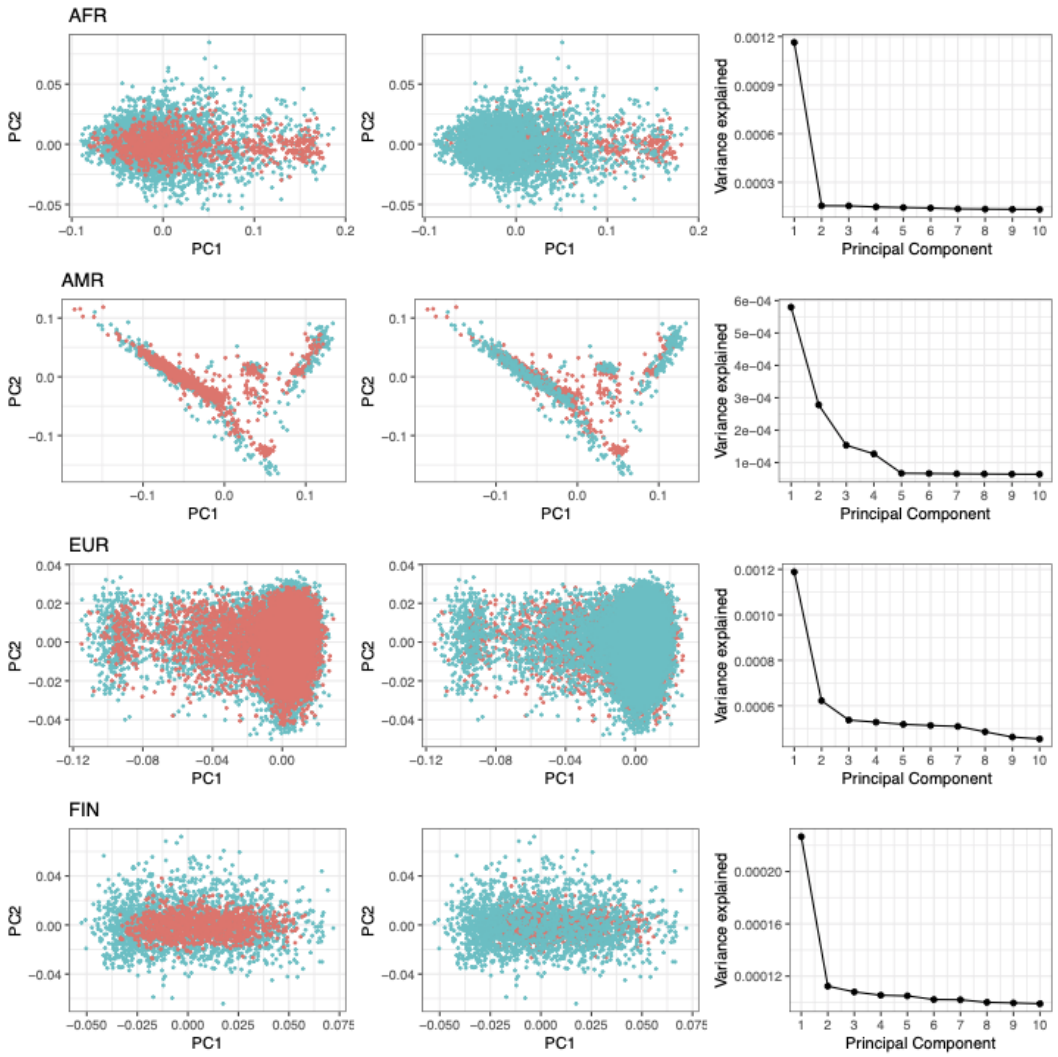

Figure 3.9: Principal component analysis for cases (red) and controls (turquoise) for each ancestry group. Cases are plotted on top (left) or bottom (middle). Scree plots (right) suggest that linear models for genetic association include up to five principal components as covariates.

Despite controlling for population stratification by analyzing major ancestry groups separately and adjusting for within-ancestry principal components in each ancestry-specific case-control analysis, the genomic inflation factors (λ_{GC}) from the complete meta-analysis (described in Section 3.2.5) was 1.40. Since the ImmunoChip intentionally covers regions of the genome previously associated with immune-mediated disease, λ_{GC} for association with T1D across ImmunoChip variants is *a priori* anticipated to be greater than one. However, it is important to differentiate inflation

due to enrichment of true biological association from inflation due to experimental artifact, such as population stratification. Due to the non-uniform distribution of ImmunoChip variants across the genome, LD Score regression (Bulik-Sullivan et al. 2015), an approach that leverages genome-wide linkage disequilibrium patterns to determine sources of inflated test statistics in GWAS, cannot be applied to this data set. Thus, to rule out population stratification, we compared λ_{GC} from the famliy-based linkage disequilibrium test (TDT) (Spielman, McGinnis, and Ewens 1993), which is robust to population stratification, to λ_{GC} from case-control analysis of comparable statistical power. Specifically, for each ancestry group, we generated five randomly sampled case-control data sets, each containing one case and one control for each trio, which results in equivalent statistical power (McGinnis, Shifman, and Darvasi 2002). For example, in our EUR cohort, there were 4,766 trios. Thus, we subsampled, out of 13,458 EUR unrelated cases and 20,143 EUR unrelated controls, five data sets each containing 4,766 cases and 4,766 controls. After excluding the major histocompatibility complex (MHC), insulin (*INS*) and protein tyrosine phosphatase, non-receptor type 22 (*PTPN22*) regions, the λ_{GC} for the EUR family-based analysis was 1.44, while the average λ_{GC} from five randomly sampled case-control data sets of equivalent power was 1.50 (Figure 3.10). Similar results are seen when only considering directly genotyped variants (Supplementary Table 6). Together, these data suggest that the inflation in the association analysis cannot be explained by population stratification in our study cohort. Thus, we believe the observed inflation in both case-control and family-based analyses is most likely due to enrichment for true association signal in ImmunoChip regions.

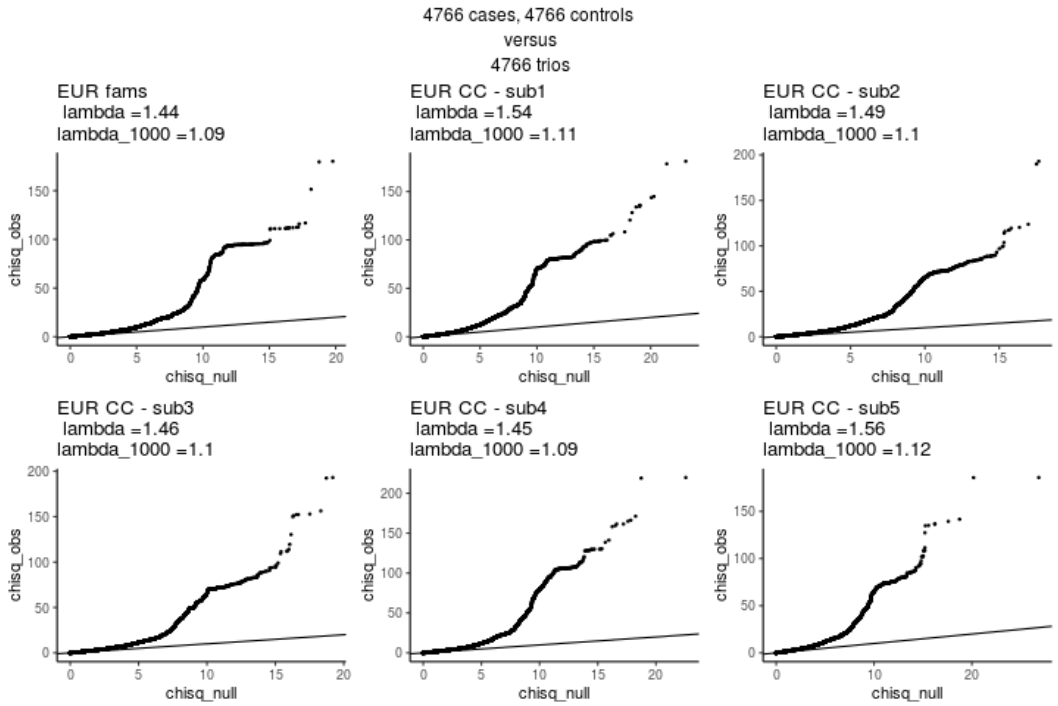

Figure 3.10: Quantile-quantile plots showing the expected chi-square association statistics against the observed chi-square association statistics from the Phase II European family-based analysis results compared to five randomly sampled European case-control cohorts with equivalent statistical power to the family-based analysis.

3.2.3 Imputation to TOPMed and 1000 Genomes reference panels

Genotypes were imputed across the entirety of all autosomal chromosomes, with the NHLBI Trans-Omics for Precision Medicine (TOPMed) Freeze 5 (Taliun et al. 2021) and 1000 Genomes phase 3 reference panels using the Michigan Imputation Server, which applied Eagle version 2.4 (Loh et al. 2016) for phasing and Minimac4 for imputation (Das et al. 2016). For each reference panel, ImmunoChip variants were aligned to the appropriate strands and reference alleles using available tools (https://www.well.ox.ac.uk/~wrayner/tools/).

Benchmarking imputation accuracy and coverage

In a subset of T1DGC samples, including 1,411 AFR, 641 AMR, and 95 EUR subjects, whole genome sequencing (WGS) data generated by the McDonnell Genome Institute at Washington University in St. Louis as part of the Centers for Common Disease Genomics (CCDG) (https://ccdg.rutgers.edu) supported by the NHGRI Genome Sequencing Program (GSP) (http://gsp-hg.org). Samples were sequenced on the Illumina HiSeq X, and sequence alignment and variant calling was performed as outlined in the standardized CCDG pipeline (Regier et al. 2018) (https://github.com/CCDG/Pipeline-Standardization/blob/master/PipelineStandard.md).

As a measure of imputation accuracy, for each single nucleotide variants (SNV) we calculated the Pearson correlation coefficient and R-squared between genotypes obtained through imputation versus WGS. To measure imputation coverage of ImmunoChip regions, we calculated the proportion of SNVs with MAF > 0.005 detected through WGS that were included in the imputed variant set after quality filtering at a range of imputation R-squared thresholds. Relative coverage of imputation based on TOPMed and 1000 Genomes reference panels was assessed in the AFR and AMR groups, for which we had adequate number of samples with available WGS. To quantify the coverage of ImmunoChip regions after imputation, we calculated the proportion of SNVs detected in WGS that were imputed with high confidence.

Imputed variant filtering

Since imputation quality (R-squared) is dependent on allele frequency and linkage disequilibrium patterns in the target population, imputed variants were filtered for ancestry-specific imputation quality (imputation R-squared > 0.8; SNPTEST info score > 0.8 in cases, controls or overall, (Marchini and Howie 2010)) and allele frequency (MAF > 0.005), and, for family-based association analyses, Mendelian inconsistency rates (< 0.01 in informative trios and parent-offspring pairs). In addition,

variants with a difference in SNPTEST info score > 0.05 between cases and controls were removed since this could artificially generate an association that is reflecting imputation differences rather than genuine differences in allele frequencies between cases and controls. Finally, only imputed variants lying within the 188 "ImmunoChip regions" (Supplementary Table 2) or in other densely genotyped regions outside of the ImmunoChip regions defined in Section 3.2.4 (Supplementary Table 3) were analyzed for association with T1D, since genotyping outside these regions on the ImmunoChip is sparse and therefore imputed variant calls less certain. However, all variants that were directly genotyped and passed QC on the ImmunoChip were included in the association analysis.

3.2.4 Defining targeted regions for discovery and fine-mapping analysis

The ImmunoChip densely covered genetic variation in the immune-associated genomic regions. Discovery analyses included all genotyped variants as well as imputed variants from any 500-kb region that contained more than 50 genotyped variants (Supplementary Table 3). To define boundaries for fine-mapping regions, we mapped previously defined ImmunoChip regions (provided by the R package humarray) from GRCh36 to GRCh38 coordinates (Supplementary Table 2): for each region, we mapped all variants originally included in the region to GRCh38 to define boundaries as the lowest and highest observed GRCh38 positions among these variants ($\pm$50 kb on either side). Fine-mapping analyses were then restricted to densely genotyped regions overlapping these ImmunoChip regions.

3.2.5 Association analysis

Phase I: Case-control analyses

All genotyped variants and high-confidence imputed variants were analyzed for association with T1D (Supplementary Tables 2 and 3). Association analyses were performed separately in the five ancestry groups. Assuming an additive mode of inheritance, we used logistic regression for unrelated case-control analyses, adjusting for five ancestry-specific principal components and using genotype posterior probabilities to account for uncertainty in the imputed genotypes using the SNPTEST version 2.5.4 software (Marchini and Howie 2010). Due to the small sample size (38 cases and 106 controls), the EAS individuals were excluded. We combined results using an inverse-variance weighted fixed-effects meta-analysis (METAL software version released on 25 March 2011)(Willer, Li, and Abecasis 2010). Forward stepwise logistic regression was performed to identify loci with more than one independent association with T1D. All conditionally independent associations ($p < 5 \times 10^{-8}$) were reported. The case-control analyses were performed under recessive and dominant models of inheritance. To evaluate the relative fit of the three models, we compared the Akaike Information Criterion (AIC) in the EUR ancestry group and identified the model providing the lowest AIC (best fit). Only genotyped variants were examined for their association with T1D on the X chromosome. The Y chromosome was not examined.

Phase II: Trio families and combined analyses

Trio families (two parents and an affected offspring) were analyzed within an ancestry group using the TDT (Spielman, McGinnis, and Ewens 1993). As TDT test statistics are susceptible to substantial bias when applied to imputed genotypes (Taub et al. 2012), a stringent variant filter was applied to the imputed genotypes, removing all variants with Mendelian inconsistencies in $> 1\%$ of trios with heterozygous offspring

or parent-offspring pairs with homozygous offspring. From the transmission disequilibrium test summary statistics, we derived effect sizes and standard-error estimates (Kazeem and Farrall 2005) and meta-analyzed with the Phase I results.

3.2.6 Statistical fine mapping

Two complementary approaches were used to define credible variant sets within each T1D-associated ImmunoChip region. Fine mapping included high-confidence variants within 750 kb of the lead variant (1.5-Mb region in total), usually consisting of imputed variants across the entire ImmunoChip region and genotyped variants adjacent to the ImmunoChip region.

Fine mapping using EUR case-control data only

Because forward stepwise model selection can fail to identify complex genetic architectures (Asimit et al. 2019), we also applied a Bayesian method (GUESSFM) in the EUR case-control data to identify the most likely combinations of variants explaining T1D risk (Wallace et al. 2015; Bottolo and Richardson 2010). An overview of the GUESSFM fine-mapping procedure is provided in Section 3.1.3 and shown in Figure 3.5. Groups of variants prioritized by GUESSFM are called "credible sets" and variants within these groups are called "credible variants." Variants that failed the quality-control metrics (or were not genotyped or imputed in our data for other reasons) but were in linkage disequilibrium ($r^2 > 0.9$ in 1000 Genomes Phase 3) with a prioritized variant were included in the comprehensive list of credible variants (Supplementary Table 11).

Trans-ancestry fine mapping

In regions where association signals were marginally associated ($p < 5 \times 10^{-4}$) in multiple ancestry groups and evidence from EUR fine mapping with GUESSFM sug-

gested a single causal variant (marginal posterior probability for one causal variant in the region > 0.5), we applied the multi-ancestry fine-mapping method PAINTOR (Kichaev and Pasaniuc 2015) to refine the association. PAINTOR uses association z-scores and population-level linkage disequilibriumto identify the combination of alleles that best explain the phenotype, multiplying the posterior probability of the causal vector across ancestry groups, assuming the same variant(s) are causal in each ancestry group. Given that the loci examined were those with evidence of one causal variant in the region, we restricted the maximum model size to two variants in the region and enumerated the posterior of every model, rather than performing a Markov-chain-Monte-Carlo search. The association z-scores used for each ancestry group were from a meta-analysis of case-controls and family trios in that ancestry cluster. For analysis with PAINTOR, linkage disequilibrium reference panels were generated with imputed genotype data from unrelated cases and controls, separately for each ancestry group, using LDstore version 1.1 (Benner et al. 2017).

3.2.7 Haplotype analyses

Haplotype analyses were performed in cases and controls of EUR ancestry by taking "best-guess" genotype values for the variants included in the analysis and obtaining haplotype phase-distribution estimates for each individual using an expectation-maximization algorithm (Excoffier and Slatkin 1995). The haplotype of each individual was sampled ten times and a logistic regression was fitted estimating the effect size of the haplotype relative to the most common haplotype in the population, with T1D status as the outcome and adjusting for five principal components. The estimates and standard errors for each haplotype relative to the most common were averaged over the ten logistic regression models to obtain the overall haplotype effect sizes on T1D risk.

3.2.8 Annotating T1D-associated protein-altering variants

The functional impacts of T1D credible variants (Supplementary Table 11) were annotated using ANNOVAR (version released on 16 April 2018, Wang, Li, and Hakonarson 2010, https://annovar.openbioinformatics.org) and the Ensembl (**ensembl.org**) and refGene (**www.ncbi.nlm.nih.gov/refseq**) annotation databases.

3.2.9 Statistical analyses

Unless otherwise noted, all statistical analyses and data visualization were performed using R version 3.6 (R Core Team 2017). All statistical tests based on symmetrically distributed test statistics were two-sided. No repeated measures data were analyzed in this study. All genotyped samples analyzed in the association tests represent distinct individuals. The R packages ggplot2 (Wickham 2016, https://ggplot2.tidyverse.org/), and cowplot (Wilke 2020, https://cran.r-project.org/web/packages/cowplot/index.html) were used for data visualization. Code used to generate the results presented in this chapter is available at https://github.com/ccrobertson/t1d-immunochip-2020.

3.3 Results

3.3.1 Genotyping and imputation of immune-related regions

After quality filtering, 61,427 participants (Figure 3.7, Supplementary Table 1) and 140,333 genotyped ImmunoChip variants were included in analyses, providing dense coverage in 188 autosomal regions ("ImmunoChip regions") (Cortes et al. 2011) and sparse genotyping in other regions (Supplementary Tables 2 and 3). Each participant was assigned to one of five ancestry groups using principal component analysis (Figure 3.8): European (EUR, $n = 47,319$), African admixed (AFR, $n = 4,290$), Finnish

(FIN, $n = 6,991$), East Asian (EAS, $n = 588$) and other Admixed (AMR, $n = 2,239$). Association analyses included 16,159 T1D cases, 25,386 controls and 6,143 trio families (i.e., an affected child and both parents) (Figure 3.9), Supplementary Tables 4 and 5). Genotypes at additional variants were imputed using the Trans-Omics for Precision Medicine (TOPMed) (Taliun et al. 2021) multi-ethnic reference panel to improve discovery and fine-mapping resolution. After imputation, the number of variants in ImmunoChip regions with imputation R-squared > 0.8 and minor allele frequency (MAF) > 0.005 in each ancestry group was 166,274 (EUR), 322,084 (AFR), 163,612 (FIN), 137,730 (EAS), and 188,550 (AMR).

Accuracy and coverage of imputation was assessed using WGS in 1,411 AFR, 641 AMR, and 95 EUR subjects. After filtering for imputation R-squared > 0.8, more than 99% of imputed SNVs within ImmunoChip regions were concordant with WGS with true R-squared > 0.5 (Figure 3.11). Among 1,411 AFR and 641 AMR subjects, 92.3% and 87.6% of variants in ImmunoChip regions detected in WGS with MAF > 0.005 were imputed with imputation R-squared > 0.8, respectively (Figure 3.12). Only variants within ImmunoChip regions or regions with relatively high variant density, defined as more than 50 variants genotyped in a 500kb region (Supplementary Tables 2 and 3), were included in the analysis, since the imputation of variants outside these regions would be based on a small number of genotyped variants only.

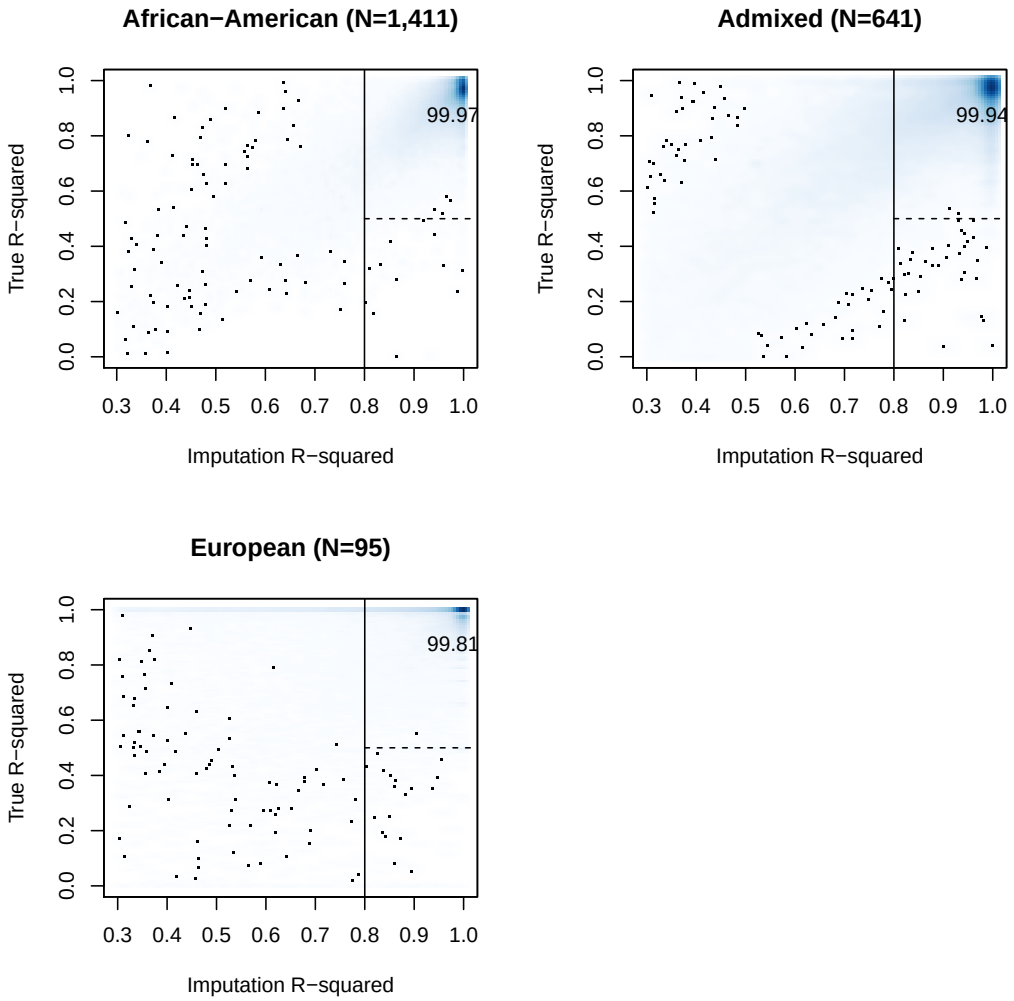

Figure 3.11: Genotype accuracy for variants in ImmunoChip regions based on a subset of 2,147 participants with available whole genome sequence (WGS) data. "Imputation R-squared" is estimated imputation quality returned by the imputation software Minimac4. "True R-squared" is the Pearson correlation between genotypes obtained through imputation to the TOPMed reference panel versus WGS. Among variants with imputation R-squared > 0.8 (right of solid vertical line), more than 90% have true R-squared > 0.5 in all three ancestry groups.

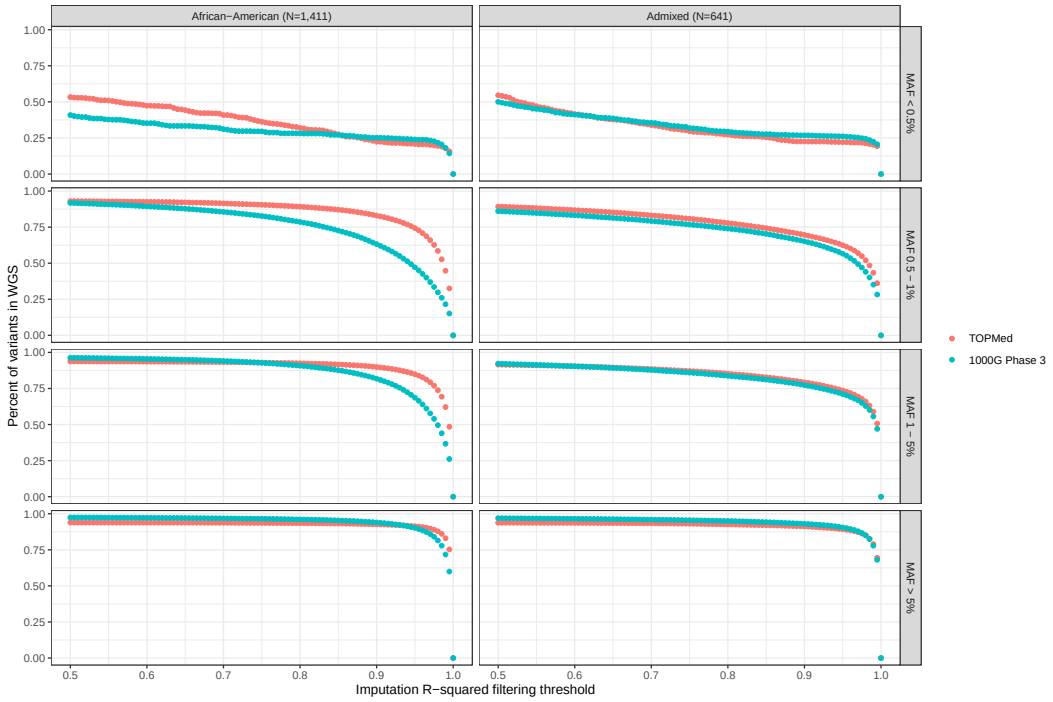

Figure 3.12: Imputation coverage of ImmunoChip regions across a spectrum of imputation quality filtering thresholds and minor allele frequencies. Y-axis shows the proportion of variants detected by whole genome sequence (WGS) data that were imputed using the TOPMed (red) or 1000 Genomes Project Phase 3 (blue) reference panel. "Imputation R-squared" is estimated imputation quality returned by the imputation software Minimac4. MAF, minor allele frequency.

3.3.2 Thirty-six new genome-wide significant regions

Initially, we analyzed unrelated cases and controls ($n = 41,545$), assuming an additive inheritance model. With minimal evidence of artificial inflation of association statistics due to population structure (Figure 3.10, and Supplementary Table 6), we identified 64 T1D-associated regions outside the major histocompatibility complex (MHC), including 24 regions associated with T1D at genome-wide significance ($p < 5 \times 10^{-8}$) for the first time. Following conditional analysis, 78 independent associations were identified ($p < 5 \times 10^{-8}$; Supplementary Table 7). On the X chromosome, the most T1D-associated variant was rs4326559 (A>C, C allele $OR = 1.09$, $p = 4.5 \times 10^{-7}$).

We extended the discovery analysis to incorporate T1D trio families ($n = 6,143$

trios, some trio families were multiplex and analyzed as multiple trios). Meta-analysis of case-control and trio results identified 78 chromosome regions associated with T1D ($p < 5 \times 10^{-8}$), including 42/43 chromosome regions previously identified in an ImmunoChip-based study (Onengut-Gumuscu et al. 2015) (rs4849135 (G>T) was $p = 2.93 \times 10^{-7}$). When comparing these 78 regions to previous T1D studies (Todd et al. 2007; Barrett et al. 2009; Onengut-Gumuscu et al. 2015; Wellcome Trust Case Control Consortium 2007; Cooper et al. 2008; Hakonarson et al. 2008; Grant et al. 2009; Bradfield et al. 2011; Huang et al. 2012; Zhu et al. 2019), 36 novel regions associated with T1D at genome-wide significance for the first time (Table 1). In the remaining 42 regions, the lead variant was within 250 kb of the lead variant in a previous T1D study. The 1q21.3 region, which contains the gene encoding the interleukin-6 receptor (IL-6R), was among the regions associated with T1D at genome-wide significance for the first time. The lead variant in this region was rs2229238 (T>C) ($p = 3.02 \times 10^{-9}$), not the nonsynonymous variant rs2228145 (A>C) (NP_000556.1:p.Asp358Ala) ($p = 2.20 \times 10^{-4}$), which was previously suggested to be causal for T1D in targeted analysis (Ferreira et al. 2013) and remains a candidate causal variant for rheumatoid arthritis (Okada et al. 2014).

Chromosome	Position (bp)[a]	Lead variant rsID	A1	A2	Putative candidate gene[b]	AF_{EUR} (A2)	OR_{meta}[c]	P_{meta}	Traits with shared association[d]
1	63643100	rs2269241	T	C	*PGM1*	0.196	1.111	4.67×10^{-12}	
1	92358141	rs34090353	G	C	*RPAP2*	0.361	1.078	1.10×10^{-8}	
1	119895261	rs2641348	A	G	*NOTCH2*	0.107	1.113	1.61×10^{-8}	Crohn's disease, T2D
1	154465420	rs2229238	T	C	*IL6R*	0.813	0.896	1.38×10^{-12}	
1	172746562	rs78037977	A	G	*FASLG*	0.124	0.884	2.41×10^{-9}	Asthma, vitiligo, allergic sensitization
1	192570207	rs2816313	G	A	*RGS1*	0.719	1.090	4.57×10^{-9}	
1	212796238	rs11120029	G	T	*TATDN3*	0.147	1.102	1.82×10^{-8}	
2	12512805	rs10169963	C	T	AC096559.1	0.580	1.074	2.78×10^{-8}	
2	100147438	rs12712067	G	T	*AFF3*	0.358	0.925	4.12×10^{-9}	
2	191105394	rs7582694	C	G	*STAT4*	0.773	0.916	2.83×10^{-9}	SLE, hypothyroidism, celiac disease, RA
2	241468331	rs10933559	A	G	*FARP2*	0.208	1.109	2.39×10^{-11}	
4	973543	rs113881148	C	A	*TMEM175*	0.626	1.082	5.72×10^{-9}	Body-fat percentage
4	38602849	rs337637	G	A	*KLF3*	0.364	0.919	2.57×10^{-10}	White blood-cell count
5	40521603	rs1876142	G	T	*PTGER4*	0.658	0.905	2.18×10^{-14}	
5	56146422	rs10213692	T	C	*ANKRD55/IL6ST*	0.241	0.912	2.85×10^{-9}	RA, Crohn's disease, MS
6	424915	rs9405661	C	A	*IRF4*	0.514	1.080	2.26×10^{-9}	
6	137682468	rs12665429	T	C	*TNFAIP3*	0.370	0.907	1.36×10^{-13}	
6	159049210	rs212408	G	T	*TAGAP*	0.638	1.112	1.42×10^{-15}	MS, Crohn's disease, eczema
7	20557306	rs17143056	A	G	*ABCB5*	0.183	0.909	2.44×10^{-8}	
7	28102567	rs10245867	G	T	*JAZF1*	0.331	0.928	3.15×10^{-8}	Eczema, hay fever, MS, SLE, monocyte percentage
8	11877675	rs2250903	G	T	*CTSB*	0.283	0.905	1.35×10^{-10}	
9	99823263	rs1405209	T	C	*NR4A3*	0.375	1.075	3.45×10^{-8}	
10	33137219	rs722988	T	C	*NRP1*	0.367	1.108	3.21×10^{-15}	
11	35267496	rs11033048	C	T	*SLC1A2*	0.366	1.091	1.53×10^{-10}	Vitiligo
11	60961822	rs79538630	G	T	*CD5/CD6*	0.035	1.213	1.14×10^{-9}	
11	61828092	rs968567	C	T	*FADS2*	0.177	0.903	8.42×10^{-9}	RA, neutrophil percentage
11	64367826	rs645078	A	C	*CCDC88B*	0.385	0.925	3.34×10^{-9}	
11	128734337	rs605093	G	T	*FLI1*	0.470	1.077	4.25×10^{-9}	
12	8942630	rs1805731	T	C	*M6PR*	0.389	1.073	4.16×10^{-8}	Eosinophil count
12	53077434	rs7313065	C	A	*ITGB7*	0.162	1.101	3.28×10^{-9}	
13	42343795	rs74537115	C	T	*AKAP11*	0.141	1.109	5.41×10^{-9}	
14	68286876	rs911263	C	T	*RAD51B*	0.710	1.083	1.69×10^{-8}	PBC, SLE, RA
16	20331769	rs4238595	T	C	*UMOD*	0.687	0.912	2.43×10^{-11}	
17	45996523	rs1052553	A	G	*MAPT*	0.232	0.879	1.65×10^{-15}	Parkinson's disease
17	47956725	rs2597169	A	G	*PRR15L*	0.348	1.081	3.35×10^{-9}	
21	44204668	rs56178904	C	T	*ICOSLG*	0.187	0.898	6.48×10^{-11}	

Of these 36 regions, 13 had a lead variant that was in strong linkage disequilibrium ($r^2 > 0.95$ in the 1000 Genomes Project European population) with variants that are associated with at least one related trait. [a]Genome build 38. [b]Closest gene or gene with mechanistic support from the literature. [c]Additive OR for the addition of an A2 allele. [d]Related traits (https://genetics.opentargets.org) where the lead variant is in strong LD ($r^2 > 0.95$ in the 1000 Genomes Project European population) with T1D lead variant. RA, rheumatoid arthritis; T2D, type 2 diabetes; SLE, systemic lupus erythematosus; MS, multiple sclerosis; IBD, inflammatory bowel disease; PBC, primary biliary cholangitis.

Figure 3.13: Newly identified regions of association with T1D with genome-wide significance ($p < 5 \times 10^{-8}$)

3.3.3 Additional regions identified using alternative inheritance models and metric of statistical significance

Applying the Benjamini-Yekutieli false discovery rate (FDR) < 0.01 (Benjamini and Yekutieli 2001) to assess statistical significance, 143 regions were associated with T1D (Supplementary Table 8). Their lead variants overlapped substantially with lead variants for 14 immune-mediated diseases from published studies, but the direction of effects frequently differed between traits (Figure 3.14). Associated variants with FDR < 0.01 but not meeting genome-wide significance ($p < 5 \times 10^{-8}$) had smaller absolute effect sizes but similar MAFs to those satisfying genome-wide significance (median (IQR) $OR = 1.07(1.06, 1.09)$ vs. $1.11(1.09, 1.13)$; median (IQR) MAF $= 0.301(0.152, 0.397)$ vs. $0.306(0.184, 0.374)$). These results indicate that remaining regions associated with T1D may have increasingly smaller effect sizes (Figure 3.15), requiring genome-wide coverage and larger sample sizes for detection.

One exception underscores the need for inclusion of understudied populations to enhance biological insight, even with limited sample sizes. On chromosome 1p22.1 near the Metal Response Element Binding Transcription Factor 2 (*MTF2*) gene, rs190514104 (G>A)) had a large effect on T1D risk (OR (95% CI) $= 2.9(1.9 - 4.5)$; $p = 6.6 \times 10^{-7}$) in the AFR ancestry group. The minor allele (A) at rs190514104 (G>A) was common in the AFR ancestry group ($> 1\%$) but rare in the others ($< 0.1\%$). Considering the limited sample size, potential heterogeneity of the AFR cohort, and possible over-estimation of effect sizes due to "the winner's curse," this association requires replication in an independent cohort. Nonetheless, this finding suggests the potential value of considering alternative metrics for defining statistical significance in genetic studies (Crouch et al. 2021).

Use of recessive and dominant models of inheritance identified 35 regions (25 dominant, 10 recessive) with a better fit than the additive model (lower AIC in Europeans) at FDR < 0.01, including nine regions that did not reach FDR < 0.01

under the additive model (Supplementary Table 9). Thus, a total of 152 regions were associated with T1D at FDR < 0.01, 143 under an additive model and nine under recessive or dominant models.

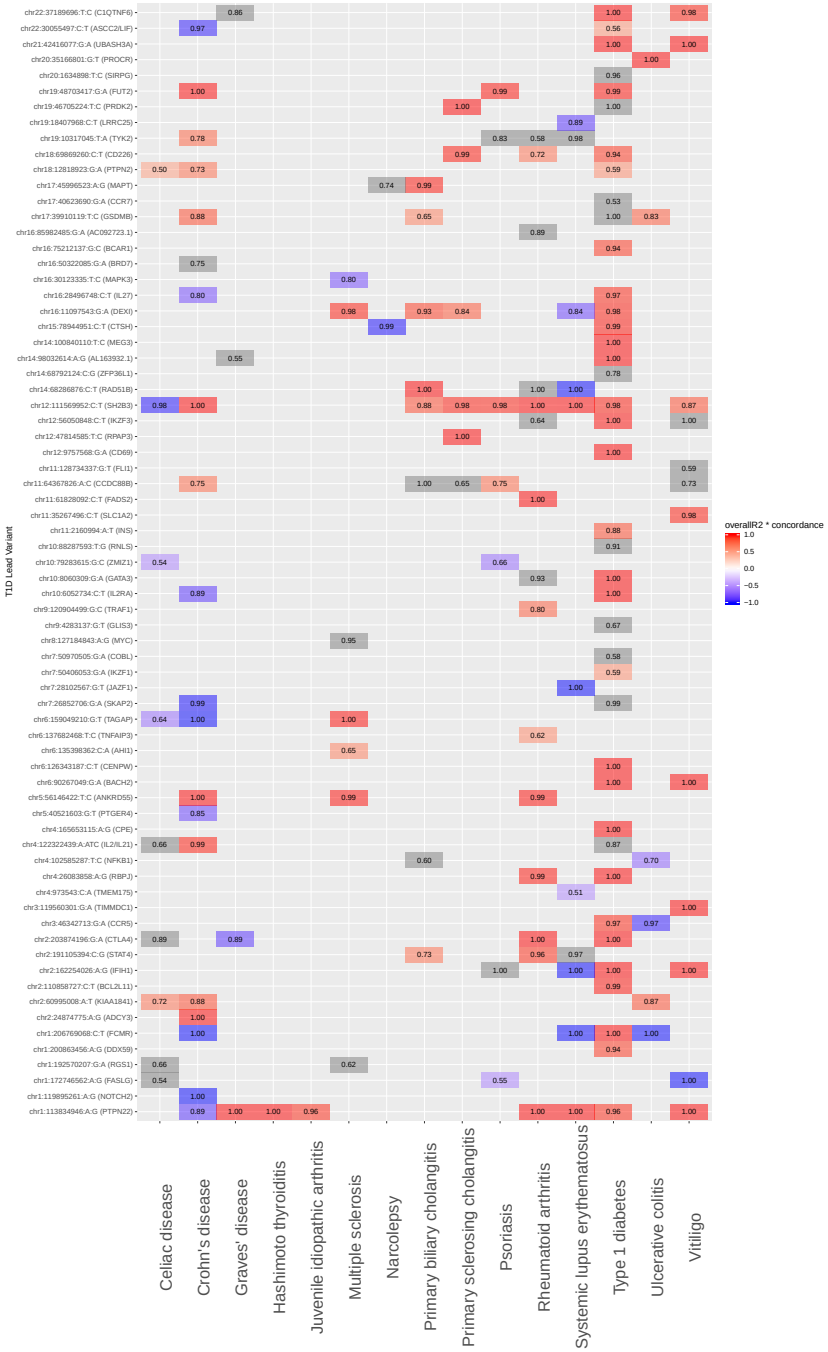

Figure 3.14: Evidence supporting shared effects between T1D and 14 immune-related diseases. 14 immune-related diseases are on the x-axis; T1D lead variants, and their corresponding candidate genes, are indicated on the y-axis. A square indicates that the corresponding disease has a genome-wide significant association in the region, with a lead variant in moderate to high linkage disequilibrium ($r^2 > 0.5$) with the lead T1D variant from this study. The r^2 between lead variants is provided within the square. Red squares indicate concordant direction of effect. Blue squares indicate discordant direction of effect. Grey squares indicate that summary statistics for T1D association with the immune-related disease lead variant was not available from this study.

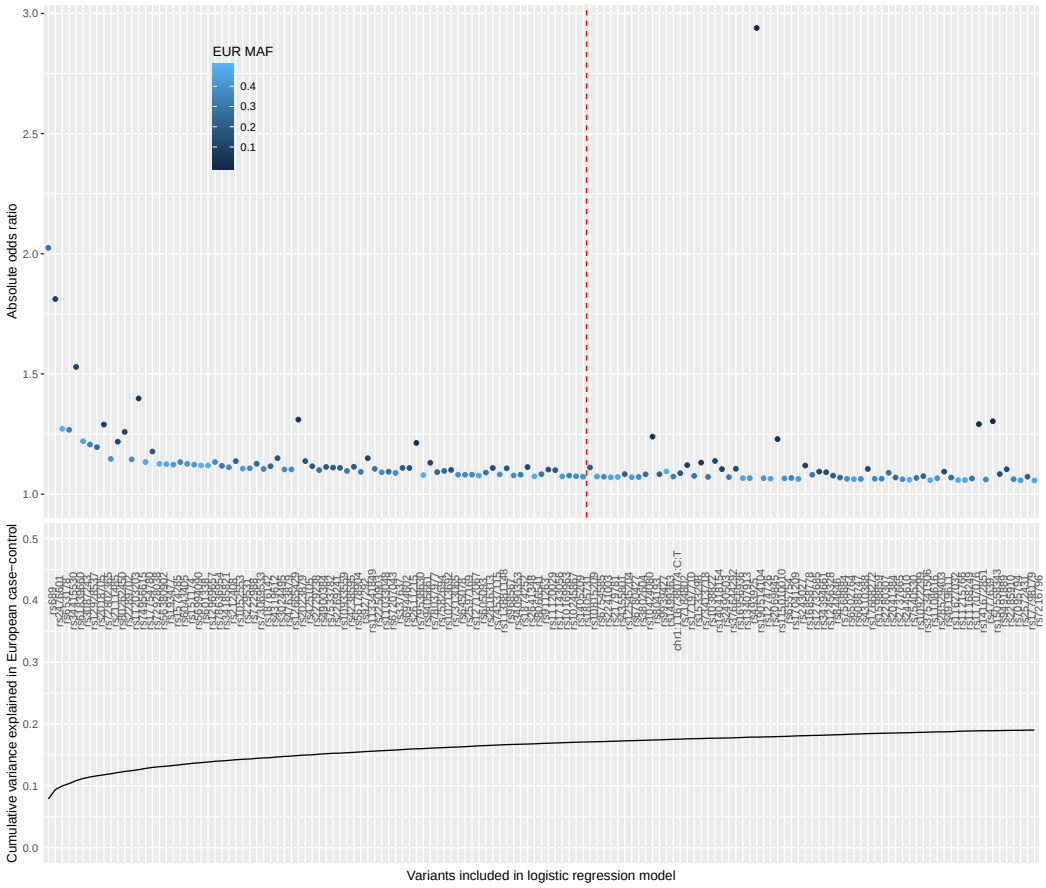

Figure 3.15: Top panel: Absolute odds ratios for the lead variant in each T1D-associated region based on FDR < 0.01. Variants are coloured by minor allele frequency (MAF) in the European ancestry collection (lighter blue corresponding to higher MAF). Those to the left of the dashed line attained genome-wide significance ($p < 5 \times 10^{-8}$). Bottom panel: Variance explained from logistic regression model using EUR case-control data only, from left to right, cumulatively adding variants to the logistic regression model; calculating the McFaddon's r^2 as a proxy for variance explained.

3.3.4 Fine mapping reveals over a third of T1D loci contain more than one independent association

To define the local architecture of T1D regions, we applied a Bayesian stochastic search method (GUESSFM, Wallace et al. 2015) to the European ancestry case-control data. Of 52 ImmunoChip regions (Supplementary Table 2) associated with T1D, GUESSFM predicted 21 (40%) to contain more than one causal variant (Fig-

ure 3.16a), compared to nine regions using stepwise conditional regression. In four regions, the lead variant in the discovery analysis was not prioritized by fine mapping (posterior probability < 0.5): 2q33.2 (*CTLA4*), 4q27 (*IL2*), 14q32.2 (*MEG3*) and 21q22.3 (*UBASH3A*). In these regions, the lead variant likely tags two or more T1D-associated haplotypes that can be identified using GUESSFM but not stepwise logistic regression, a phenomenon observed previously (Wallace et al. 2015; Asimit et al. 2019). For example, although stepwise regression analysis in the *UBASH3A* locus supported a single causal variant (Supplementary Table 7), GUESSFM fine mapping and haplotype analyses indicated that the lead variant in this region, rs11203203 (G>A), is unlikely to be causal. GUESSFM fine mapping supported a three-variant model (rs9984852 (T>C), rs13048049 (G>A), and rs7276555 (T>C)) (Figure 3.16b), which had a better fit than the single variant model (AIC 45073 vs. 45138, Figure 3.16c). Haplotype analysis demonstrated that when rs11203203 (G>A) is present without the GUESSFM-prioritized variants, there is no effect of rs11203203 (G>A) on T1D risk (Figure 3.16d). Resampling experiments consistently supported two or more causal variants in the region, with at least one of the three GUESSFM-prioritized variants more likely to be causal than rs11203203 (G>A) (Supplementary Table 10). Given the complexity of association in the *UBASH3A* region, and likely at many loci, statistical methods designed to use univariable summary statistics alone are not sufficient to explore the genetic architecture of T1D. We provide the comprehensive list of T1D credible variants and haplotype analyses for all 52 fine-mapped regions (Supplementary Table 11, https://github.com/ccrobertson/t1d-immunochip-2020).

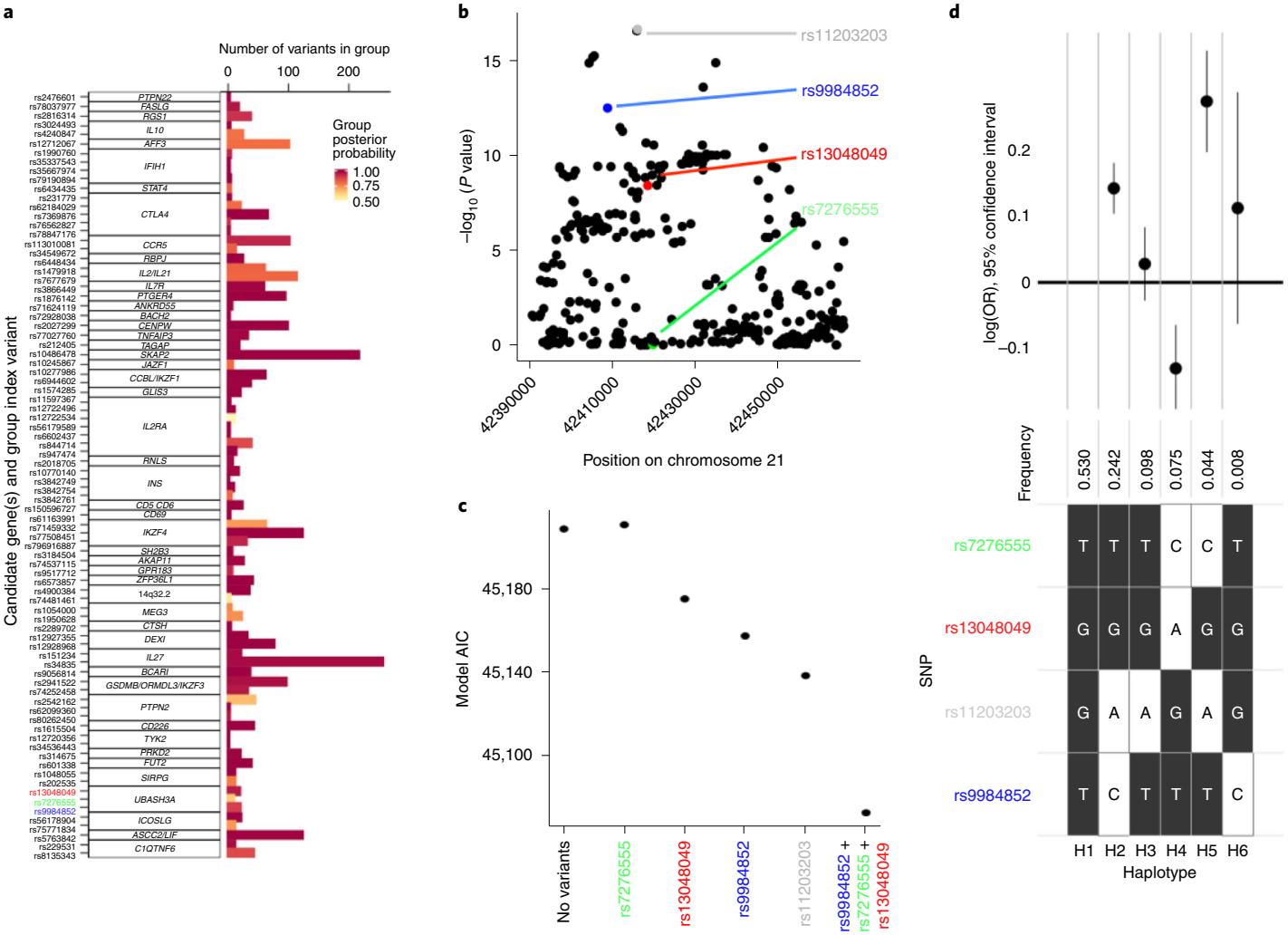
a
Number of variants in group
0
100
200
Group posterior probability
1.00
0.75
0.50
Candidate gene(s) and group index variant
PTPN22
FASLG
RGS1
IL10
AFF3
IFIH1
STAT4
CTLA4
CCR5
RBPJ
IL2/IL21
IL7R
PTGER4
ANKRD55
BACH2
CENPW
TNFAIP3
TAGAP
SKAP2
JAZF1
CCBL/IKZF1
GLIS3
IL2RA
RNLS
INS
CD5 CD6
CD69
IKZF4
SH2B3
AKAP11
GPR183
ZFP36L1
14q32.2
MEG3
CTSH
DEXI
IL27
BCARI
GSDMB/ORMDL3/IKZF3
PTPN2
CD226
TYK2
PRKD2
FUT2
SIRPG
UBASH3A
ICOSLG
ASCC2/LIF
C1QTNF6
b
rs11203203
rs9984852
rs13048049
rs7276555
−log10 (P value)
15
10
5
0
42390000
42410000
42430000
42450000
Position on chromosome 21
c
Model AIC
45,180
45,140
45,100
No variants
rs7276555
rs13048049
rs9984852
rs11203203
rs9984852 + rs7276555 + rs13048049
d
log(OR), 95% confidence interval
0.2
0.1
0
−0.1
Frequency
0.530
0.242
0.098
0.075
0.044
0.008
SNP
rs7276555
rs13048049
rs11203203
rs9984852
T T T C C T
G G G A G G
G A A G A G
T C T T T C
H1 H2 H3 H4 H5 H6
Haplotype

Figure 3.16: Fine mapping of T1D regions using a Bayesian stochastic search algorithm. a, Number of variants in GUESSFM-prioritized groups with group posterior probability > 0.5. The candidate gene names and lead variants for each group are shown on the y-axis. b, Manhattan plot of the *UBASH3A* region from the EUR case-control analysis highlighting the lead variant from the univariable analysis rs11203203 (G>A) (gray) and the three variants prioritized using GUESSFM - rs9984852 (T>C) (blue), rs13048049 (G>A) (red) and rs7276555 (T>C) (green). c, Comparison of model AIC in the *UBASH3A* region for models fit using EUR cases and controls only, comparing combinations of alleles prioritized either in univariable (gray) or GUESSFM analyses (red, green and blue). d, Analysis of haplotypes associated with T1D in the *UBASH3A* region. The most common haplotype (H1: T-G-G-T for rs7276555-rs13048049-rs11203203-rs9984852) is presented on the far left; alternative haplotypes (H2-H6) are shown with white squares highlighting the differentiating alleles (C, A, A or C, respectively). The frequency and effect estimates for association with T1D relative to the baseline haplotype (H1) are shown above the grid (the point and error bars represent the log-transformed OR and 95% confidence interval of the log-transformed OR, respectively); for example, the log-transformed OR for T1D risk for haplotype H3 (T-G-A-T) relative to the baseline haplotype (H1) is close to zero and the 95% confidence interval crosses zero. Haplotype analyses were performed based on $n = 33,601$ unrelated EUR individuals (13,458 T1D cases and 20,143 controls).

3.3.5 Multi-ethnic fine mapping further refines credible sets in 4p15.2, 6q22.32 and 18q22.2

Differences in linkage disequilibrium between ancestry groups can be advantageous in prioritizing causal variants (Wojcik et al. 2019a). In the 30 regions where analysis suggested a single causal variant, we performed multi-ethnic fine-mapping using PAINTOR (Kichaev and Pasaniuc 2015). Eight regions identified an associated variant ($p < 5 \times 10^{-4}$) in more than one ancestry group: five with associations in EUR and FIN, and three with associations in EUR and AFR. In three regions, the number of variants prioritized was markedly reduced by including multiple ancestry groups: 4p15.2 (*RBPJ*), 6q22.32 (*CENPW*) and 18q22.2 (*CD226*) (Figures 3.17a, 3.18, and 3.19, and Supplementary Table 12). In the chromosome 4p15.2 (*RBPJ*) region, the credible set from EUR ancestry contained 24 variants. In contrast, using

PAINTOR with EUR and AFR summary statistics, only five variants were prioritized with a posterior probability > 0.1 (Figure 3.17a). Among these prioritized variants, rs34185821 (A>G) and rs35944082 (A>G), both located in the non-coding transcript *LINC02357*, have the potential to disrupt multiple transcription factor binding motifs (Boyle et al. 2012). rs35944082 (A>G) also overlaps open chromatin in multiple adaptive immune cell types (Figure 3.17b) and resides in a FANTOM enhancer site (Lizio et al. 2015). Further, rs34185821 (A>G) is one of three prioritized variants flanking an activation-dependent Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) peak in lymphocytes and a stable response element in human islets35, with potential to perturb an extended TATA box motif (Ward and Kellis 2016).

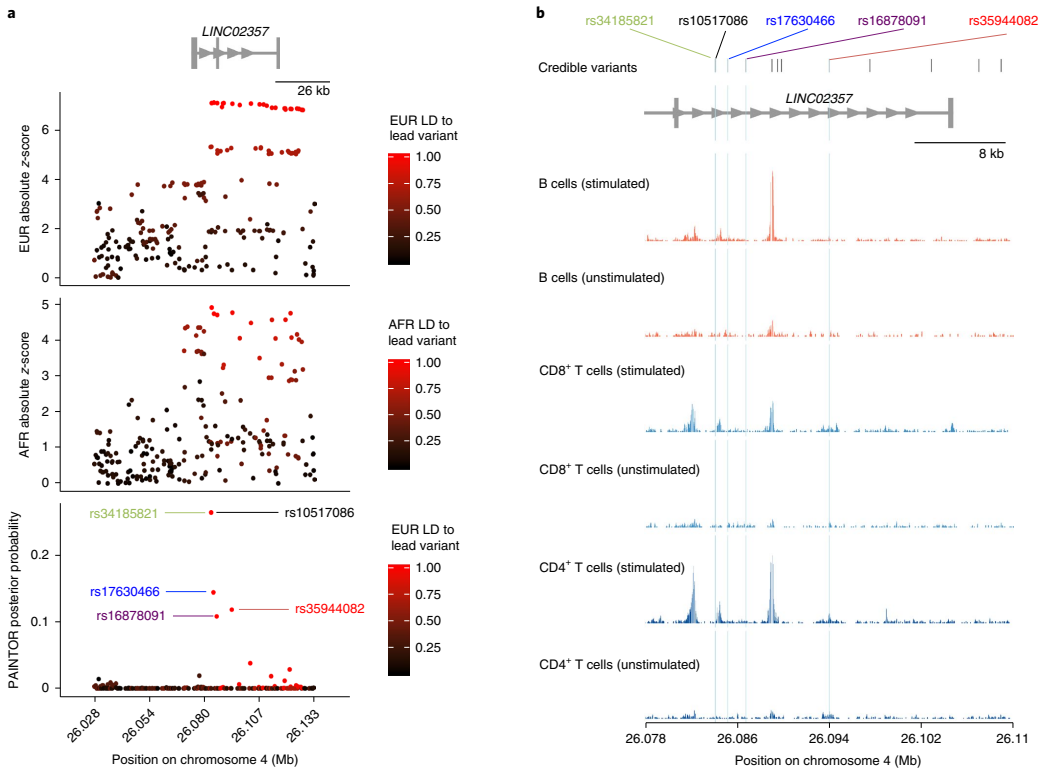

Figure 3.17: Fine mapping of the chromosome 4p15.2 region. a, Association z-score statistics for the EUR (top) and AFR (middle) ancestry groups; posterior probabilities (bottom) from multi-ancestry fine mapping of the EUR and AFR groups using PAINTOR. The z-scores are colored according to the linkage disequilibrium value to the lead PAINTOR-prioritized variant. b, Overlay of T1D credible variants with open chromatin ATAC-seq peaks in immune cells, with the variants prioritized by PAINTOR (posterior probability > 0.1) indicated with blue dashed lines. The normalized ATAC-seq read count is shown for stimulated and unstimulated $CD4^+$ T cells, $CD8^+$ T cells and B cells.

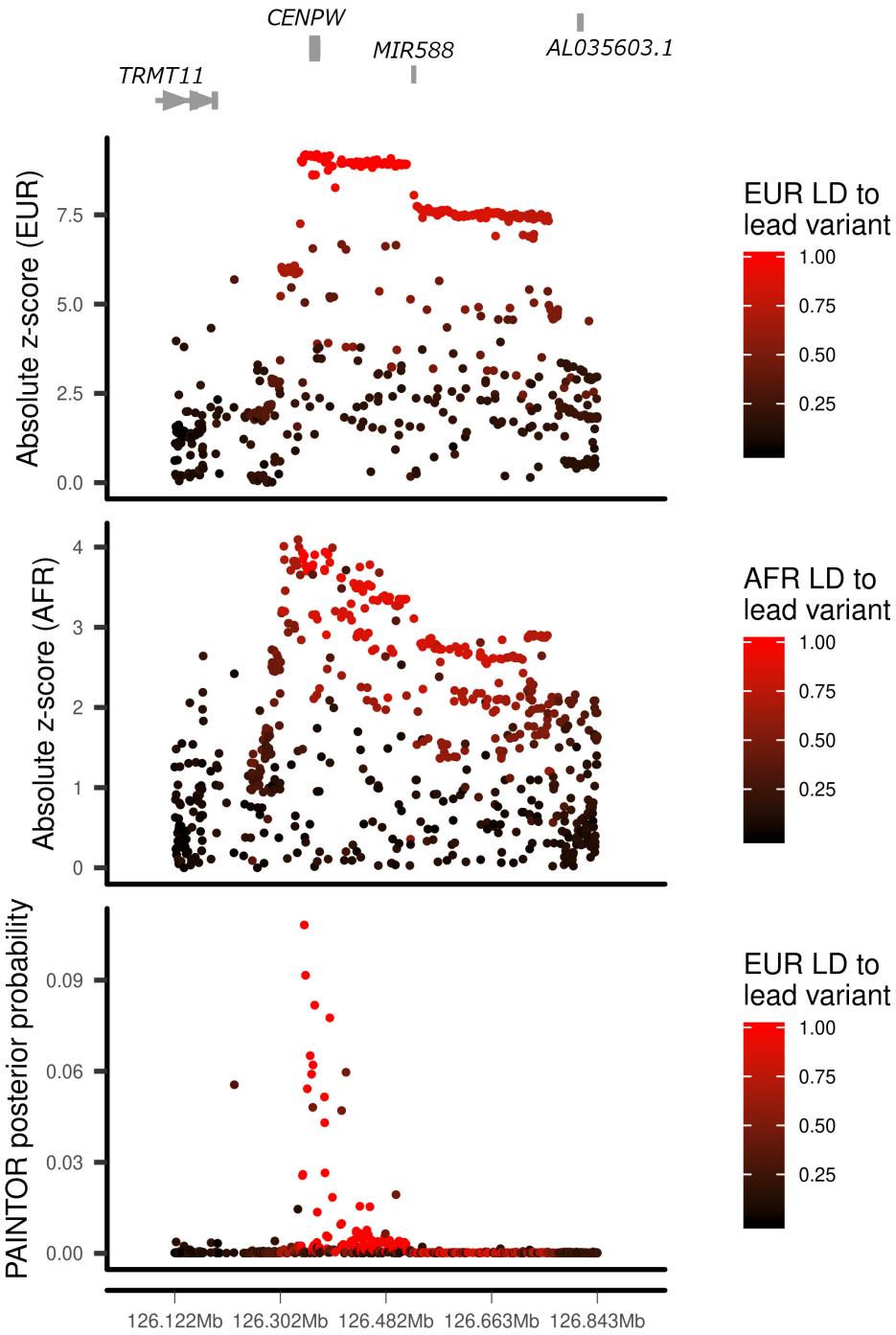

Figure 3.18: Fine mapping of the chromosome 6q22.32 region. European (EUR, top panel) and African (AFR, middle panel) ancestry group association z-score statistics and posterior probabilities (bottom panel) from multi-ethnic fine mapping of EUR and AFR using PAINTOR. z-scores are colored by linkage disequilibrium (LD) to the lead PAINTOR-prioritized variant.

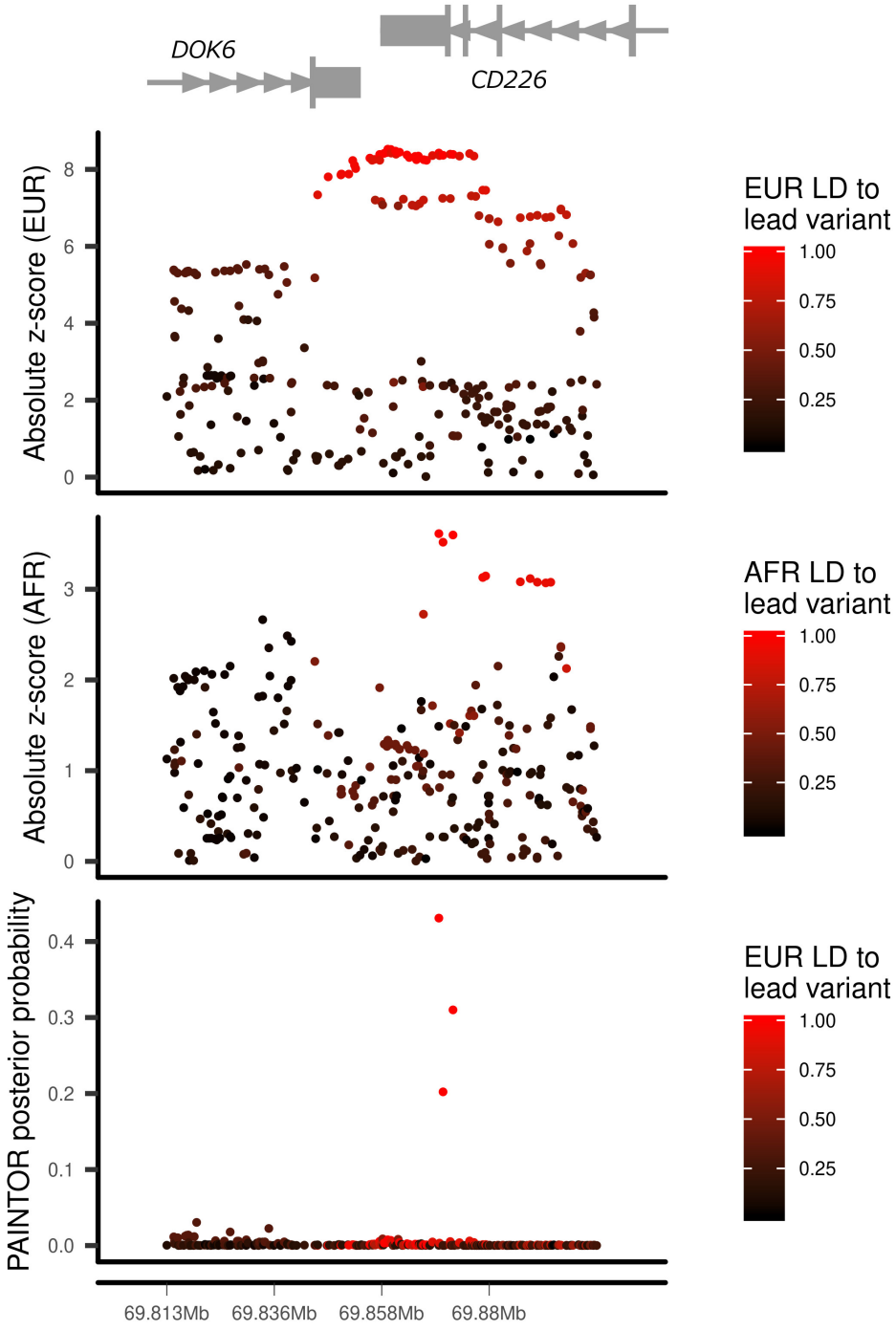

Figure 3.19: Fine mapping of the chromosome 18q22.2 region. European (EUR, top panel) and African (AFR, middle panel) ancestry group association z-score statistics and posterior probabilities (bottom panel) from multi-ethnic fine mapping of EUR and AFR using PAINTOR. z-scores are colored by linkage disequilibrium (LD) to the lead PAINTOR-prioritized variant.

3.3.6 T1D-associated protein-altering variants

Only 34 of 2,732 (1.2%) credible variants (group posterior probability > 0.5) were protein-altering (nonsynonymous, frameshift, stop-gain, or splice-altering) with 12 having support for a role in T1D (Supplementary Table 13). We identified several previously unreported protein-altering variants as highly prioritized in T1D credible sets (posterior probability > 0.1): a protective missense variant in *UBASH3A*, rs13048049 (G>A) (p.Arg324Gln; OR = 0.84; AF_{EUR} = 0.051); two low-frequency splice donor variants in *IFIH1*, rs35732034 (C>T); (OR = 0.63; AF_{EUR} = 0.0089) and rs35337543 (C>G) (OR = 0.61; AF_{EUR} = 0.0099); and a missense variant in *CTLA4*, rs231775 (A>G) (p.Thr17Ala; OR = 1.20; AF_{EUR} = 0.36).

3.4 Discussion

In the largest genetic analysis of T1D to date, we identified 36 novel regions at genome-wide significance and implicated a total of 152 regions outside the MHC in T1D susceptibility at FDR < 0.01. We refined the set of putative causal variants and number of independent associations in many T1D regions through increased sample size, dense genotyping and imputation, inclusion of diverse ancestry groups, and optimized analytical approaches to fine mapping.

Existing models of the genetic basis of complex traits suggest that disease-associated genetic variation tends to follow a predictable pattern, where common variants have small effects on disease risk, while variants with more potent effects are less common due to selective pressures (Figure 3.20).

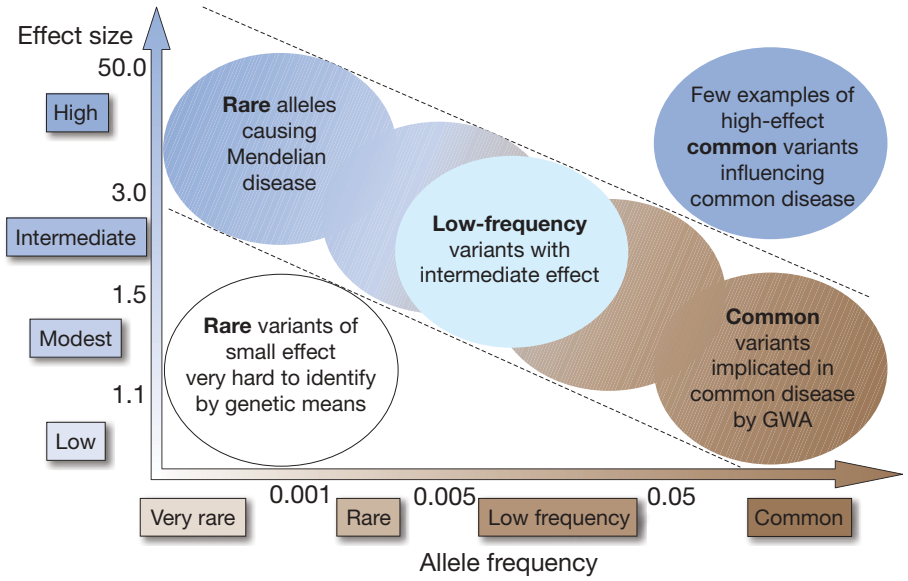

Figure 3.20: Feasibility of identifying genetic variants by risk allele frequency (x-axis) and strength of genetic effect (y-axis, odds ratio). Figure and caption from Manolio et al. 2009.

With some notable exceptions, including genetic variation in the HLA and insulin regions, genetic associations with T1D are largely consistent with this model (Figure 3.21). In novel regions (red points in Figure 3.21), lead variants were typically common alleles with modest effects on disease risk. Since the vast majority of common variants in one ancestry group are also present at some frequency in all continental ancestry groups (1000 Genomes Project Consortium 2015), it is unlikely that there remain unidentified common variants with large effect on T1D risk. However, there may still be many low-frequency or rare variants with large effects on T1D risk, particularly in non-European ancestry populations. This possibility was underscored by the association results from our analysis of only 1,045 African-ancestry T1D cases (Figure 3.22), where a novel risk variant, rs190514104 (G>A) near *MTF2*, had an unusually large effect on T1D risk (OR= 2.9). The effect size in this locus is likely overestimated due to winner's curse, especially since the sample size remains small. We also acknowledge the possibility that association is confounded by admixture due to the heterogeneity of AFR subjects. Nonetheless, this finding highlights the po-

tential value of focusing future T1D genetic studies in previously underrepresented populations.

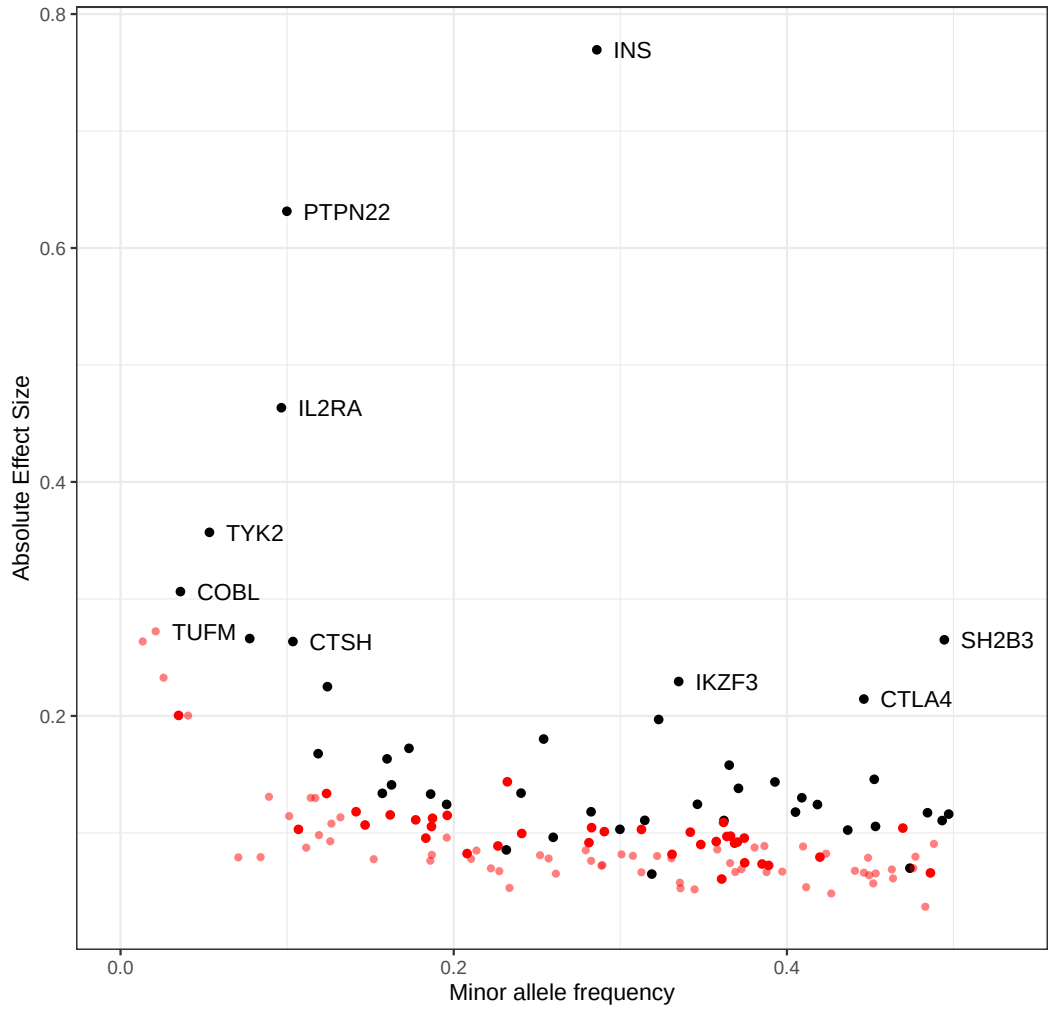

Figure 3.21: European-ancestry allele frequency and effect size distribution for variants associated with type 1 diabetes (T1D). Each point represents a lead variant from one of the 152 T1D-associated regions from our meta-analysis. Variants are positioned according to their absolute effect size (y-axis = log(OR)) and minor allele frequency (x-axis) in the EUR case-control analysis. red, associated with T1D at genome-wide significance ($p < 5x10^{-8}$) for the first time; light-red, associated with T1D at FDR< 0.01 but not $p < 5x10^{-8}$; black, previously associated with T1D at genome-wide significance.

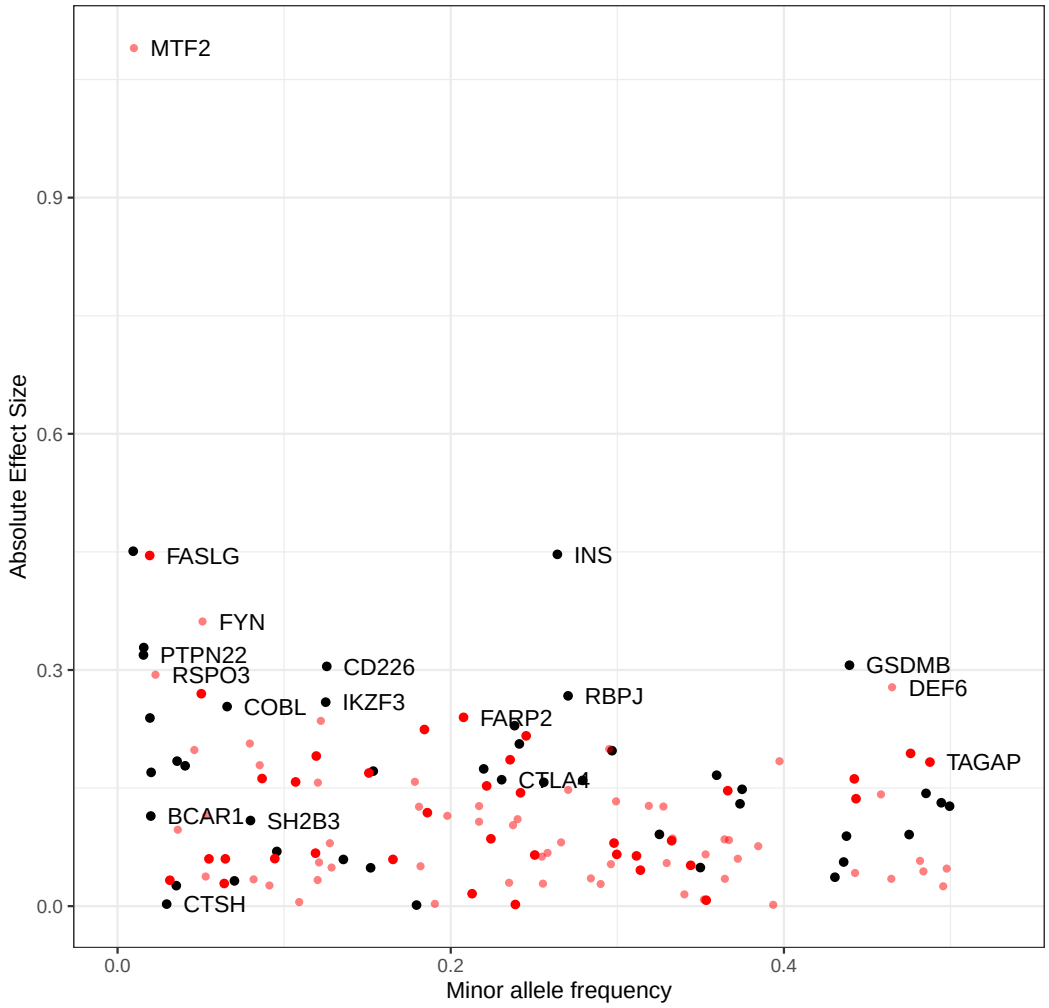

Figure 3.22: African-ancestry allele frequency and effect size distribution for variants associated with type 1 diabetes (T1D). Each point represents a lead variant from one of the 152 T1D-associated regions from our meta-analysis. Variants are positioned according to their absolute effect size (y-axis = log(OR)) and minor allele frequency (x-axis) in the AFR case-control analysis. red, associated with T1D at genome-wide significance ($p < 5x10^{-8}$) for the first time; light-red, associated with T1D at FDR$<$ 0.01 but not $p < 5x10^{-8}$; black, previously associated with T1D at genome-wide significance.

Considering genetic variants on the ImmunoChip are restricted to regions that cover less than 5% of the genome, it is impressive that 152 regions in this analysis show evidence of association with T1D risk. This reflects the extensive pleiotropy between immune-mediated diseases, and in particular, the shared genetic basis of T1D and

other seropositive autoimmune diseases. Interestingly, in some regions with evidence for shared causal variants between T1D and other immune-mediated diseases, alleles protective against T1D may increase risk of another disease (or the other way around, Figure 3.14). This occurred most frequently with Crohn's disease. Out of nineteen regions where evidence supported a common causal variant in T1D and Crohn's disease, nine (47%) had discordant direction of effect (lead variants near *PTPN22*, *NOTCH2*, *FCMR*, *PTGER3*, *TAGAP*, *SKAP2*, *IL2RA*, *IL27*, *ASCC2/LIF*). In contrast, for rheumatoid arthritis, all ten regions sharing a likely causal variant with T1D had concordant directions of effect (lead variants near *PTPN22*, *STAT4*, *CTLA4*, *RBPJ*, *ANKRD55*, *TNFAIP3*, *TRAF1*, *FADS2*, *SH3B3*, *CD226*). Functional enrichment analyses have shown substantial enrichment of Crohn's disease-associated variants in myeloid-specific accessible chromatin regions, while variants associated with T1D, and other seropositive autoimmune diseases, primarily show enrichment in lymphoid-specific accessible regions (Chiou et al. 2021). Thus, pleiotropic variants with discordant effects on T1D and Crohn's disease may influence the balance or interactions between myeloid- and lymphoid-mediated immune responses.

Fine-mapping revealed that over a third of T1D-associated regions have multiple causal variants, including several regions harboring established immune regulators (e.g., *IL10*, *IFIH1*, *CTLA4*, *CCR5*, *IL2*, *IL2RA*, *TYK2*, and *UBASH3A*). This trend of allelic heterogeneity located near immune regulatory factors raises the possibility that a substantial portion of T1D risk may be mediated by the constellation of variants and haplotypes shaping expression of these genes, which converge on a common set of immune regulatory pathways.

One limitation of this study is that genotyping was restricted to ImmunoChip content, which provides dense coverage in 188 immune-relevant genomic regions, as defined by previous largely EUR ancestry-based GWAS of immune-related traits. This design restricts the scope of discovery and fine mapping, and generalizability

of subsequent functional enrichment analyses. Although this analysis is the largest and most comprehensive study of T1D genetics, extension of future genetic studies to genome-wide analyses (Crouch et al. 2021; Chiou et al. 2021) and continuing efforts to expand cohorts from diverse populations will further define the genetic landscape of T1D.

Chapter 4

Functional prioritization of type 1 diabetes-associated variants with chromatin accessibility profiles

This chapter is adapted from:

> **Robertson CC**, Inshaw JRJ, Onengut-Gumuscu S, Chen WM, Flores Santa Cruz D, Yang H, Cutler AJ, Crouch DJM, Farber E, Bridges SL Jr., Edberg JC, Kimberly RP, Buckner JH, Deloukas P, Divers J, Dabelea D, Lawrence JM, Marcovina S, Shah AS, Greenbaum CJ, Atkinson MA, Gregersen PK, Oksenberg JR, Pociot F, Rewers MJ, Steck AK, Dunger DB; Type 1 Diabetes Genetics Consortium; Wicker LS, Concannon P, Todd JA, and Rich SS. Fine-mapping, trans-ancestral and genomic analyses identify causal variants, cells, genes and drug targets for type 1 diabetes. *Nature Genetics*. 2021 Jun 14:1-0.

Supplementary tables referenced in this chapter can be obtained at: https://doi.org/10.1038/s41588-021-00880-5.

4.1 Background

4.1.1 Genome regulation and chromatin accessibility

Only about 1% of the human genome encodes proteins. Much of the remaining non-coding sequence is involved in regulating expression of genes in different contexts (e.g., developmental timepoints, cell types, or in response to stimuli). Understanding the precise mechanisms by which non-coding sequences regulate gene expression remains a major challenge. Several genome-wide approaches can be used to measure molecular features that correlate with regulatory function (Section 1.2.4). Integration of these diverse approaches has provided catalogues of regulatory regions (e.g., enhancers, promoters, silencers, and insulators) across human tissues and cell types (Kundaje et al. 2015; Moore et al. 2020).

In general, DNA regulatory elements regulate gene expression by recruiting transcription factors that bind specific DNA sequences, which tends to cause increased chromatin accessibility at those sites. Additionally, these sequence-specific binding events can lead to recruitment of additional proteins that remodel the surrounding chromatin leaving it more accessible for subsequent binding events (Allis and Jenuwein 2016). Therefore, one approach to assess the regulatory function of a genomic region is to determine the degree to which chromatin at that location is physically accessible. Chromatin accessibility is determined by nucleosome occupancy and the dynamics of other chromatin-associated proteins, such as transcription factors and proteins that direct chromatin organization (Klemm, Shipony, and Greenleaf 2019). Measures of chromatin accessibility can be interpreted primarily as evidence for displacement of nucleosomes by transcription factors (Figure 4.1, Thurman et al. 2012).

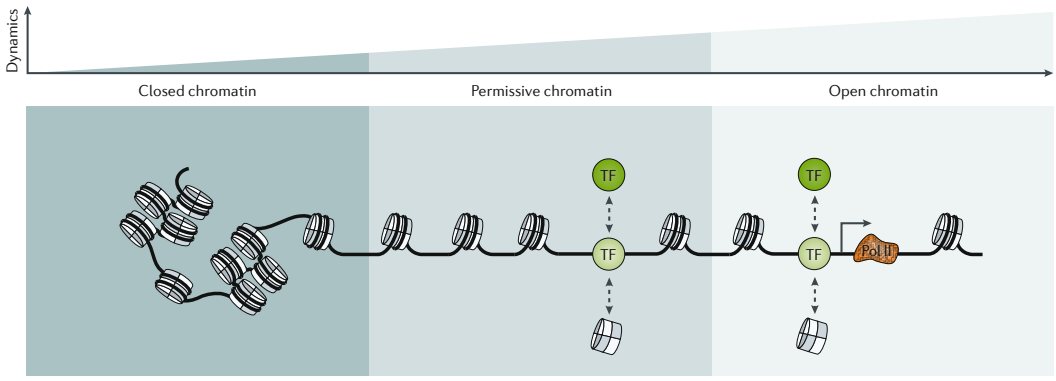

Figure 4.1: A continuum of accessibility states broadly reflects the distribution of chromatin dynamics across the genome. In contrast to closed chromatin, permissive chromatin is sufficiently dynamic for transcription factors to initiate sequence-specific accessibility remodelling and establish an open chromatin conformation (illustrated here for an active gene locus). Pol II, RNA polymerase II; TF, transcription factor. Figure and caption obtained from Klemm, Shipony, and Greenleaf 2019.

There are several well-established methods that are commonly used to profile chromatin accessibility genome-wide (Boyle et al. 2008; Schones et al. 2008; Buenrostro et al. 2013). Assay for Transposase Accessible Chromatin using sequencing (ATAC-seq) uses a hyperactive Tn5 transposase to cleave and ligate high throughput sequencing adapters to exposed chromatin (Figure 4.2). ATAC-seq is a popular approach for studying chromatin accessibility in patient samples because it requires relatively few cells per experiment compared to other approaches.

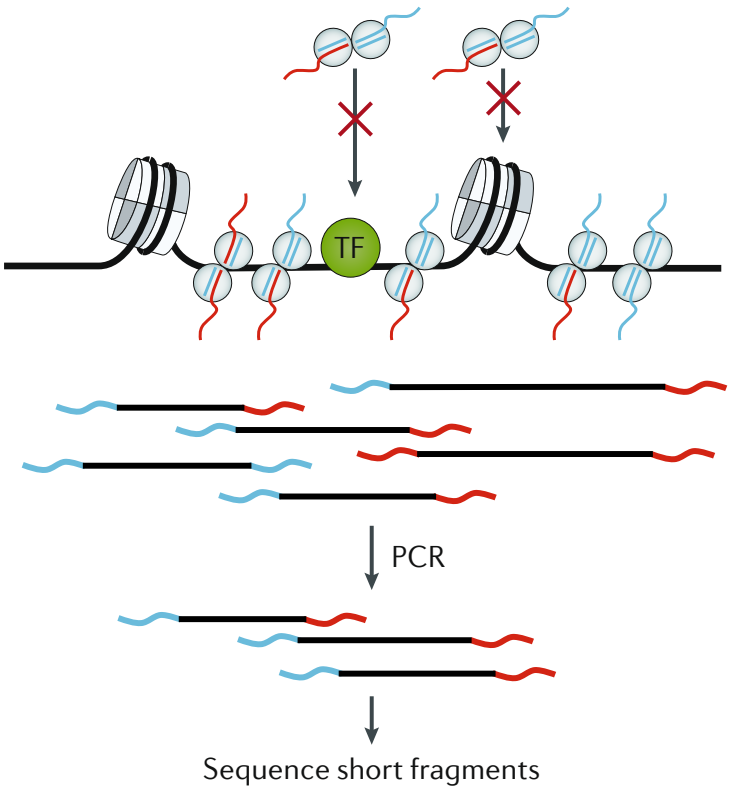

Figure 4.2: Assay for transposase-accessible chromatin using sequencing (ATAC-seq) uses a hyperactive transposase (Tn5) to simultaneously cleave and ligate adaptors to accessible DNA. Figure and caption adapted from Klemm, Shipony, and Greenleaf 2019.

4.1.2 Molecular quantitative trait loci

Mapping molecular quantitative trait loci (QTL) is an approach to identify genetic variants with regulatory function (Figure 4.3). To generate QTL maps, molecular traits (e.g., gene expression, DNA methylation, metabolite, or protein levels) and genetic variation (e.g., genotypes from a DNA microarray) are measured across many individuals (typically hundreds). Using these data, relationships between genetic variation and molecular traits can be systematically explored. QTL mapping can be performed for any molecular trait that can be robustly quantified in hundreds of individuals, including gene expression ("eQTLs"), RNA splicing ("sQTLs"), DNA

methylation ("mQTLs"), chromatin acetylation ("acQTLs") and chromatin accessibility ("caQTLs"). Genetic variants can influence nearby (*cis*-QTLs) or distant (*trans*-QTLs) molecular traits.

Large-scale efforts to map molecular effects of genetic variation have demonstrated that molecular QTLs are abundant. For example, the Genotype-Tissue Expression (GTEx) project, a large collaborative effort to map eQTLs across diverse human tissues, found a *cis*-eQTL in at least one tissue for 94.7% of protein-coding genes (GTEx Consortium 2020). Another study, which generated data on RNA expression, DNA methylation, H3K4me1 and H3K27ac in diverse immune cell types, found that, within a purified cell type, genetic variation accounts for the majority of gene expression variance across individuals (Chen et al. 2016). Analyses from GTEx suggest that QTLs tend to act either within a narrow tissue context or broadly across all tissues (GTEx Consortium 2020).

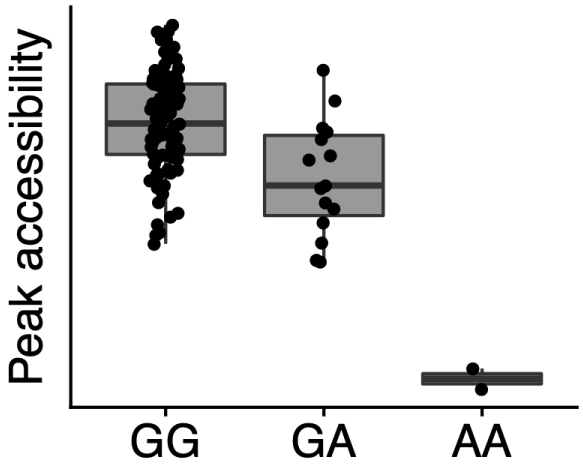

Figure 4.3: An example of a chromatin accessibility quantitative trait locus (caQTL), where the A allele at rs72928038 is associated with decreased chromatin accessibility in the genomic interval chr6:90266766-90267747. The x-axis indicates genotypes at the SNP rs72928038. The y-axis indicates the accessibility of chromatin.

4.1.3 Interpreting genetic association results using functional genomics

Unlike genetic variants causing rare Mendelian diseases, which tend to disrupt protein-coding sequences, the majority of genetic variants associated with common complex diseases are in non-coding regions of the genome, which obscures the causal genes and mechanisms underlying their association (Maurano et al. 2012). Functional genomics tools can help to interpret results from genetic association studies in several ways (Cano-Gamez and Trynka 2020).

First, maps of regulatory elements can be used to identify disease-relevant cell types. Efforts such as the ENCODE (Moore et al. 2020) and Roadmap Epigenomics (Kundaje et al. 2015) projects have provided epigenetic maps for most major cell lineages in the human body. These data sets can be used to obtain an unbiased estimation of the relative contribution of major cell types and organ systems to a given disease. Specifically, for each cell type, the genome can be annotated according to regulatory features and segmented into discrete intervals of differing states (e.g., open/closed chromatin, or inferred chromatin states (Ernst and Kellis 2017)). Then, one can determine whether disease-associated genetic variants are preferentially overlapping a particular regulatory feature compared to random distribution across the genome. These and other analyses have shown that variants identified in genetic association studies are preferentially located in accessible chromatin (Maurano et al. 2012). Furthermore, genetic variants associated with a particular disease are enriched in regulatory regions from cell types important to the underlying disease process. For example, genetic variants associated with autoimmune diseases are enriched in immune cell enhancers (Farh et al. 2015; Onengut-Gumuscu et al. 2015). In some instances, these analyses have implicated less obvious cell types in disease etiology. Functional enrichment analysis of BMI-associated genetic variants have implicated cell types from the central nervous system (Loos 2018).

Second, molecular QTLs can be used to prioritize disease-associated variants with the potential to affect gene regulation. When a disease-associated variant influences a proximal molecular trait (i.e., when a disease variant is also a QTL), hypotheses focused on regulatory mechanisms underlying the association with disease can be examined in more detail. However, there are important statistical considerations when integrating genetic association results with molecular QTLs. In particular, due to the frequency of genetic variants influencing molecular traits and extensive linkage disequilibrium in human populations, most genetic variants are *associated* with expression of at least one nearby trait in some tissue, but in most cases this is due to linkage disequilibrium, not causal biology (Liu et al. 2019). To avoid false inference of causal molecular traits (e.g., genes or regulatory elements) mediating variant-disease association, statistical "co-localization" methods can be used to formally estimated the probability that a variant is causally related to both disease and a molecular trait (Figure 4.4).

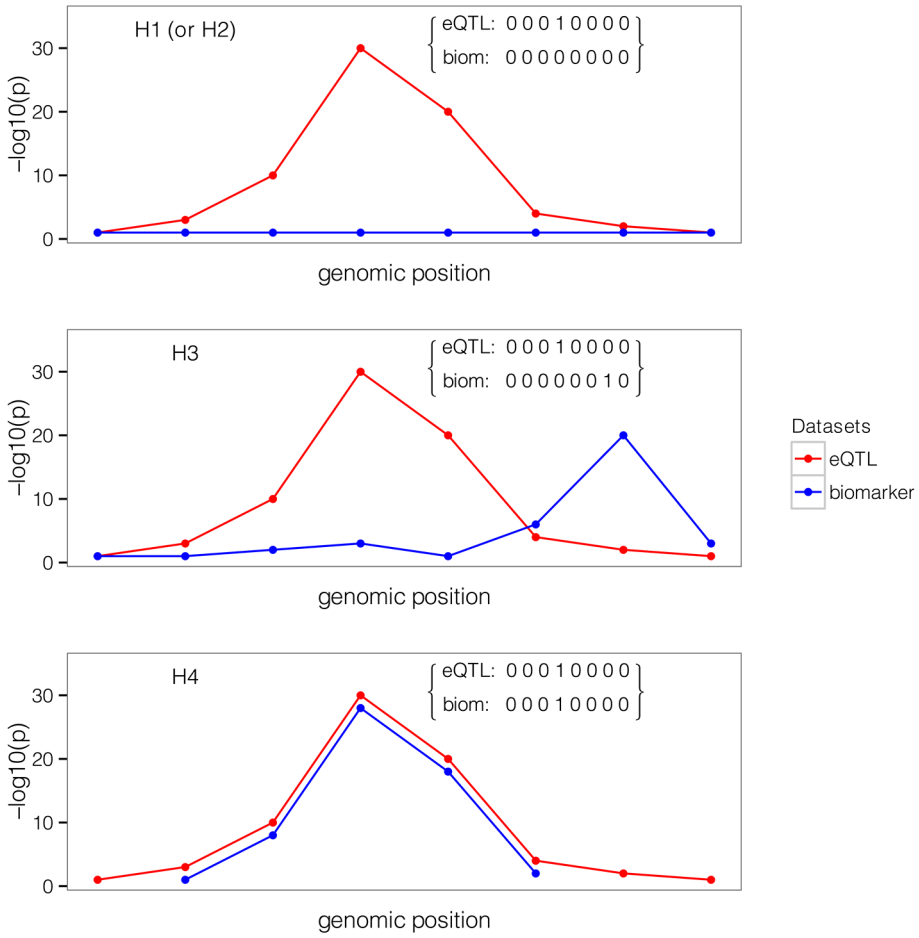

Figure 4.4: An example region showing association patterns under different hypotheses. In this example, there are 8 variants in the region and two traits of interest, a biomarker (blue) and expression of an arbitrary gene (red). For each trait, the hypothesis is represented by a binary vector. The value of 1 means that the variant is causally involved in disease, 0 means that it is not. Under H1 and H2, only one dataset shows an association. Under H3, the causal SNP is different for the biomarker dataset compared to the expression dataset. Under H4, there is a single causal variant underlying both the biomarker association and the eQTL. Figure adapted from Giambartolomei et al. 2014.

4.1.4 Motivation

Genetic screening and autoantibody surveillance can detect islet autoimmunity before overt progression to T1D (Sharp et al. 2019; Krischer et al. 2019; Onengut-Gumuscu et al. 2019), providing an opportunity for prevention. Multiple immune therapies have

been explored in clinical trials (Skyler 2018). Recently, a 14-day course of teplizumab, an anti-CD3 monoclonal antibody, delayed T1D in high genetic-risk individuals by a median of two years (Herold et al. 2019). This success shows that appropriately timed immune-modulating therapy can alter the autoimmune process preceding disease onset. Defining the genetic variants contributing to T1D risk and how they disrupt immune pathways may lead to more precise therapeutic targets, better characterization of their role in disease initiation and progression, and improved opportunities for safe and effective intervention and, ultimately, prevention of T1D (King, Davis, and Degner 2019; Nelson et al. 2015).

About half of T1D risk is genetic, with the other half driven by non-genetic factors that remain poorly defined. While many genetic risk loci have now been identified from GWAS and refined with fine mapping, the specific causal variants and their mechanisms of action for most T1D loci remain unknown. We showed previously that T1D credible variants are most strongly enriched in lymphocyte and thymic enhancers (Onengut-Gumuscu et al. 2015). Yet, resolving causal variants, mapping them to genes, and determining causal mechanisms remains a challenge. The first step in to address this challenge is to generate specific hypotheses about molecular effects of T1D variants. Here, we use chromatin accessibility quantitative trait loci (caQTLs) from $CD4^+$ T cells to prioritize credible variants for interrogation of molecular mechanisms underlying T1D association. We present a compelling hypothesis of genetic regulatory mechanism in the T1D locus encoding the transcription factor, BACH2. Finally, by integrating implicated *cis*-mechanisms (i.e., eQTL genes for T1D credible variants) with immune protein-protein interaction networks, we identify existing therapeutic drugs targeting T1D candidate genes and networks.

4.2 Methods

4.2.1 Generating representative cell-type- and condition-specific chromatin-accessibility profiles

Publicly available ATAC-seq data were obtained for diverse immune-cell types (Calderon et al. 2019), pancreatic islets (Varshney et al. 2017; Ramos-rodríguez et al. 2019) and cardiac fibroblasts (Jonsson et al. 2016).

We generated ATAC-seq data on $CD4^+$ T cells ($n = 6$ donors) and $CD19^+$ B cells ($n = 4$ donors) cells under different culture and stimulation conditions. The $CD4^+$ T cells were enriched and stimulated as previously described (Burren et al. 2017). Briefly, $CD4^+$ T cells were isolated from whole blood using RosetteSep (STEMCELL technologies, Canada) according to the manufacturer's instructions. Cells were left untreated or stimulated with Dynabeads human T activator CD3/CD28 beads (Invitrogen, UK) at a ratio of 1 bead : 3 cells for 4 h at 37 °C and 5% CO_2.

The $CD19^+$ B cells were positively selected from peripheral blood mononuclear cells using anti-CD19 beads (Miltenyi Biotec, GmbH) and cultured for 24 h in X-VIVO 15 (Lonza) supplemented with 1% human Ab serum (Sigma) and penicillin-streptomycin (Thermo Fisher), and plated in 96-well CELLSTAR U-bottomed plates (Greiner Bio-One) at a concentration of 2.5×10^5 cells/well. The cells were left untreated or stimulated with 10 μg/ml goat anti-human IgM/IgG/IgA antibody (109-006-064, Jackson Immunoresearch), 0.15 μg/ml rhCD40L (ALX-522-110-C010, ENZO Lifesciences), and 20 ng/ml rhIL-21 and rhIL-4 (200-21 and 200-04, respectively, Peprotech) for 24 h.

ATAC-seq data were generated from 50,000 cells from each cell type and culture condition following the Omni-ATAC protocol (Corces et al. 2017). The ATAC-seq datasets were mapped to GRCh38.p12 (Harrow et al. 2012) using minimap2 (version 2.17, Li 2018), except for GSE123404 (pancreatic-islets dataset), where bowtie2 (ver-

sion 2.3.5, Langmead and Salzberg 2012) was used. After mapping, the technical replicates (where available) were merged and PCR-duplicated reads were detected using Picard tools (version 2.20.2, http://broadinstitute.github.io/picard). The percentage of detected duplicated reads was very low (mean value $< 1\%$) in all datasets. Next, bigWig files were generated using bamCoverage from the deeptools package (version 3.3.0, Ramírez et al. 2016), using reads-per-genome-coverage normalization and ignoring allosomes and the mitochondrial chromosome. Peaks were called using macs2 (version 2.1.2, Gaspar 2018) with the parameters "–nomodel –shift 37 –extsize 73 –keep-dup all."

The immune-cell ATAC-seq dataset (GSE118189, Calderon et al. 2019) was used to create a consensus list of peaks. For each cell type, the donor contributing the fewest number of reads to that cell type was selected and the number of reads was divided by two. The reads were then randomly pooled by that number for each sample, creating a representative alignment file for that cell type. This procedure was performed twice to obtain two pseudo-replicates. Peaks were called using macs2 with the same parameters. The irreducible discovery rate (IDR) was calculated between the two pseudo-replicates (Li et al. 2011), any peak with an IDR ≤ 0.05 were included in the consensus list of peaks. This list was then used as a feature reference and the reads were counted per feature using featureCounts from the package subread (version 1.6.4, Liao, Smyth, and Shi 2014). A similar approach was used for the other datasets in the analysis. The IDR was used to obtain a reliable list of peaks. In these datasets, no feature reference was derived from the IDR and counting was performed directly from the list obtained from GSE118189. Workflows were implemented using conda and snakemake, and are available at https://github.com/dfloresDIL/MEGA .

4.2.2 ATAC-seq enrichment analyses

To examine the enrichment of T1D credible variants (group marginal posterior probability > 0.8 from GUESSFM) in open chromatin, two complementary approaches were used for each cell type. First, we developed a "SNP-matching" approach, in which variants were randomly sampled across the genome and matched on linkage disequilibrium (LD) structure and gene density to generate a null distribution of SNPs overlapping accessible chromatin. Specifically, the number of T1D credible variants falling within open chromatin was compared with variants in regions of the genome with similar LD structure and gene density as follows:

1. Using EUR individuals from 1000 Genomes Project data, all variants with $r^2 > 0.8$ to each other were identified.

2. The T1D credible variants with group marginal posterior probability > 0.8 were binned with regards to their LD block size: 1-9, 10-19, 20-49, 50-74,75-99,100-149 or 150-249.

3. The 1000 Genomes Project data variants were binned with regards to LD block size, taking an LD block as the variants with $r^2 > 0.8$ with an index variant.

4. For each T1D credible group, an LD block from the 1000 Genomes Project data of the same bin size and with the same (or similar for large haplotypes) number of genes overlapping the credible group was randomly selected; therefore, a similar number of variants to the T1D credible group with an approximately equivalent LD structure and gene density was selected.

5. Repeated step four 100 times, yielding 100 randomly sampled genome segments with an approximately equivalent size and LD structure to the T1D credible variants.

6. For cell type X, the number of T1D credible SNPs overlapping ATAC-seq peaks was counted. This was compared with the number overlapping ATAC-seq peaks from the first randomly sampled set of variants. The z-score (Fisher's exact test) was calculated for the comparison of ATAC-seq peak overlap with T1D credible variants versus randomly sampled variants with equivalent size, gene density and LD structure.

7. Repeated step six 100 times, one for each randomly sampled set of haplotypes across the genome, thereby obtaining 100 z-scores.

8. The mean z-score from the 100 tests was compared with a normal distribution to obtain an enrichment p-value for cell type X. Steps six and eight were performed for each cell type and condition.

Second, we applied a published method called GoShifter (Trynka et al. 2015), which tests for enrichment of trait-associated SNPs in any arbitrary set of genomic intervals (typically representing functional genomic annotations). Unlike the SNP-matching approach described above, which generates a null distribution by sampling sets of SNPs with similar features, GoShifter generates a null distribution within each locus by randomly shifting the genomic annotation sites (Figure 4.5). In our application, the genomic annotation was open chromatin peaks.

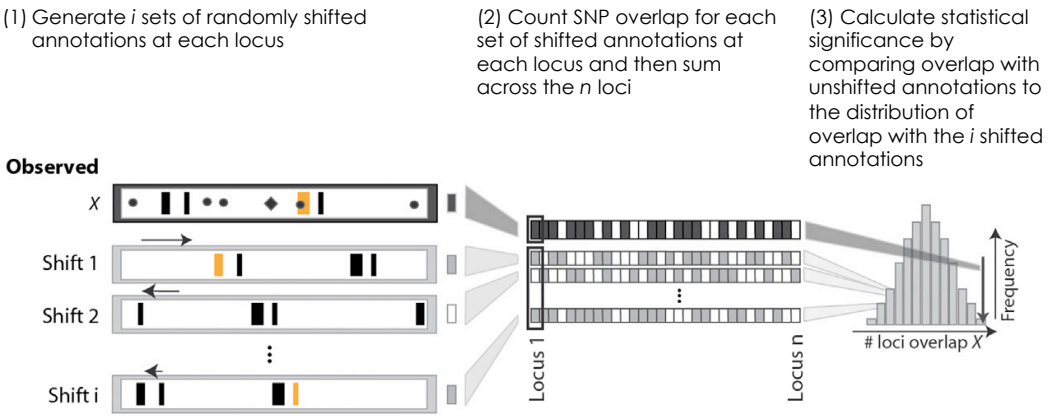

Figure 4.5: GoShifter approach to determining significance of overlap between credible SNPs and a genomic annotation. (1) We generate a null distribution by randomly shifting (blackarrows) X sites within each region i times. To ensure the same number of annotations in a region with each random shift, we circularize the region. (2) For each random shift, we count how many SNPs overlap the shifted annotations in each region, and then sum the number of SNPs overlapping shifted annotations across all regions, providing i values to generate a null distribution. (3) We estimate the significance of the observed overlap by comparing the sum of overlap with unshifted annotations to the null distribution. Figure adapted from Trynka et al. 2015.

Credible set enrichment in condition-specific accessible chromatin

For each of 24 cell types, examining only peaks in the consensus list of peaks from ATAC-seq dataset GSE118189 (25 immune cell types), we defined peaks with significantly increased accessibility (FDR < 0.01) after stimulation ("stimulation-specific peaks") and peaks with significantly decreased accessibility after stimulation ("unstimulated-specific peaks") using the R package DESeq2 (Love, Huber, and Anders 2014). We tested for enrichment of credible variants in condition-specific peaks using the SNP-matching approach described above.

4.2.3 Generating caQTL maps using T1DGC frozen samples

We profiled chromatin accessibility in 115 individuals (57 controls and 58 T1D cases; 67 AFR and 48 EUR) from the Type 1 Diabetes Genetics Consortium (T1DGC). $CD4^+$ T cells were purified from viably frozen peripheral blood mononuclear cells using magnetic cell separation according to the manufacturer's protocol, using either negative ($n = 42$; STEMCELL Technologies EasySep human $CD4^+$ T-cell isolation kit) or positive ($n = 73$; MACS Miltenyi Biotec) selection. The selection approach was incorporated in the data processing and analysis. After $CD4^+$ T-cell purification, the "Omni-ATAC-seq" protocol (Corces et al. 2017) was followed for nuclei isolation, transposase incubation and library preparation. The libraries were sequenced using 75-bp paired-end reads on an Illumina NextSeq.

Data were processed using the PEPATAC pipeline (http://pepatac.databio.org, Smith et al. 2020). Briefly, the reads were trimmed using Skewer (version0.2.2, Jiang et al. 2014) and, after removing reads mapping to mitochondrial and human repeat regions, were mapped to GRCh38 using bowtie2 (Langmead and Salzberg 2012). The PCR duplicates were removed, enzyme cut sites were inferred based on read alignment and peaks were called using macs2 (Gaspar 2018). Libraries with transcription-start-site enrichment scores below 6 or fewer than 10×10^6 aligned reads were excluded from the analyses. A set of consensus peaks was determined by merging peaks across all samples using BEDOPS (version 2.4.35, Neph et al. 2012). A matrix of peak counts was calculated by counting the number of cut sites within each consensus peak in each sample using the R package bigWig (https://github.com/andrelmartins/bigWig). Peaks with low counts were excluded (required ≥ 10 reads in $\geq 50\%$ of samples). We confirmed matching sample identity between ATAC-seq libraries and genotyped individuals using the "Match BAM to VCF" (MBV) command in the software tool set QTLtools (Fort et al. 2017).

Further peak quality filtering and normalization were performed using the R pack-

age edgeR (Robinson, Mccarthy, and Smyth 2010). These steps included:

1. filtering for peaks with $\geq$ 10 counts per million across samples within each batch,

2. peak-count normalization using the trimmed mean of M-values (TMM) method (Robinson and Oshlack 2010),

3. mean-variance modeling-based transformation using the "voom" function to enable linear modeling of peak counts assuming a normal distribution, and

4. removing outlier peaks by clustering samples based on the counts for each peak (one at a time using k-means with $k = 2$) and excluding any peak that resulted in one sample clustering separately from all of the other samples.

Association between imputed genotype dosage and chromatin accessibility (caQTL analysis) was tested using a linear model, adjusting for the first two genotype principal components, age at sample collection, transcription-start-site enrichment score and $CD4^+$ T-cell purification approach using the R package MatrixEQTL (Shabalin 2012). The caQTL discovery analyses were performed separately by ancestry group (EUR and AFR) and combined in an inverse-variance-weighted fixed effect meta-analysis (R package meta, Balduzzi, Rücker, and Schwarzer 2019). All variant-peak combinations were tested where the accessibility peak was within 1 Mb of a T1D credible variant.

4.2.4 Co-localization analysis

We evaluated co-localization of T1D and caQTL for all peaks where at least one T1D credible variant (as defined by GUESSFM) was associated with peak accessibility (meta-analysis $p < 5 \times 10^{-5}$) using the R package coloc (Giambartolomei et al. 2014) and visualized co-localized signals using the R package locuscomparer (Liu et al. 2019). Conditional summary statistics were used in regions predicted to have more

than one causal variant underlying the T1D association or regions with multiple, conditionally independent variants associated with accessibility of the same peak. When running coloc for T1D-caQTL co-localization, we used a prior probability of co-localization of 5×10^{-6} and provided association β and standard errors as input data. When running coloc for T1D-eQTL co-localization, we used the same priors and supplied association z-scores. We considered GWAS and QTL signals to be significantly co-localized when the posterior probability of co-localization was greater than 0.8 (PP.H4.abf > 0.8).

4.2.5 Allele-specific accessibility analysis

For significant caQTLs that co-localized with T1D-associated variants, we tested for allele-specific accessibility of the caQTL peak. First, we identified individuals heterozygous for T1D credible variants overlapping the caQTL peak. For each heterozygous individual, we then counted the number of reads overlapping the variant position containing the reference or alternative allele. We only performed this analysis if the T1D credible variant overlapping the caQTL peak was directly genotyped on the Immunochip, as uncertainty in the heterozygous status of an individual could lead to biased results. For peaks with at least five participants who had at least five reads overlapping the peak, we formally tested whether the proportion of reads containing an alternative allele deviated significantly from the expected null hypothesis proportion of 0.5. We calculated the p-values for deviation from "allelic balance" (proportion = 0.5 for each read) by fitting a generalized linear mixed model where the dependent variable is the number of reads and follows a Poisson distribution, and the independent variables include a fixed effect for the allele and a random effect for the participant.

4.2.6 Supershift EMSA

Jurkat cells (E6-1) were purchased from the American Type Culture Collection and cultured in RPMI-1640 medium (Gibco) supplemented with 10% fetal bovine serum, 1% penicillin-streptomycin and 1% sodium pyruvate at 37 °C and 5% CO_2.

Labeled (5' IRDye 700) and unlabeled 31-bp, single-stranded oligonucleotides containing rs72928038 were obtained from Integrated DNA Technologies (reference allele strand, 5'-AGGGACGGATTTCCTGTAAGCTGATCTTGAA-3'; and alternative allele strand, 5'-AGGGACGGATTTCCTATAAGCTGATCTTGAA -3') along with complementary oligonucleotides. Double-stranded oligonucleotides were generated by annealing equal amounts of labeled or unlabeled complementary oligonucleotides at 95 °C for 5 min, followed by gradual cooling with a ramp rate of -1.2 °C min^{-1} for 1 h (Bio-Rad C1000 Touch Thermal Cycler).

Nuclear extract from Jurkat cells was obtained by following the manufacturer's protocol for the NE-PER nuclear and cytoplasmic extraction reagents kit (Thermo Scientific) and the extracted nuclear protein was dialyzed with Slide-A-Lyzer MINI dialysis units, 10,000 MWCO (Thermo Scientific) against 1 l buffer (10 mM Tris, pH 7.5, 50 mM KCl, 200 mM NaCl, 1 mM dithiothreitol, 1 mM phenylmethylsulfonyl fluoride and 10% glycerol) for 16 h at 4 °C with slow stirring.

The binding reaction for the EMSA was carried out using 2 μl 10x binding buffer (100 mM Tris, 500 mM KCl and 10 mM dithiothreitol; pH 7.5), 2 μl of 25 mM dithiothreitol (2.5% Tween 20), 1 μl poly(dI-dC) (1 μg μl^{-1} in 10 mM Tris and 1 mM EDTA; pH 7.5), 1 μl of 1% NP-40, 100 mM $MgCl_2$, 20 fmol IRDye double-stranded oligonucleotide probe and 16 μg Jurkat nuclear extract in a final volume of 20 μl. For the supershift lanes, tested transcription-factor-binding antibodies (ETS1 rabbit mAb and Stat1 rabbit mAb) were diluted 1:50 with ddH_2O. Negative-control rabbit IgG was diluted to the same concentration as the tested antibody. Diluted antibody (1 μl) was added to the binding reaction mixture while maintaining a total volume

of 20 μl. The binding reaction was incubated for 20 min at room temperature, after which 2 μl 10x Orange loading dye was added to the reaction. Electrophoresis was performed with binding reaction mixture on a pre-run 6% DNA retardation gel for 70 min at 70 V. To capture the image, the gel was placed directly on the Odyssey-CLx (Licor) scan bed. The gel was scanned with a thickness of 0.5 mm in the 700-nm channel. The EMSA binding condition for rs72928038 was repeated three times to ensure reproducibility of the experiment.

4.2.7 Priority index

To prioritize drug targets implicated by T1D genetic associations, we ran the priority-index algorithm, as implemented in the R package Pi (Fang et al. 2019). Data used to identify eQTL co-localization (eGenes) included

- unstimulated monocytes ($n = 414$), lipopolysaccharide-stimulated monocytes after 2 h ($n = 261$), lipopolysaccharide-stimulated monocytes after 24 h ($n = 322$), interferon-γ-stimulated monocytes after 24 h ($n = 367$) from Fairfax et al. 2014;
- unstimulated B cells ($n = 286$) from Fairfax et al. 2012;
- unstimulated natural killer cells ($n = 245$) (unpublished);
- unstimulated neutrophils ($n = 114$) from Andiappan et al. 2015;
- unstimulated $CD4^+$ T cells ($n = 293$) and unstimulated $CD8^+$ T cells ($n = 283$) from Kasela et al. 2017;
- whole blood97 ($n = 5,311$) from Westra et al. 2013
- whole-blood meta-eQTL-analysis ($n = 31,684$) from Võsa et al. 2018.

Hi-C data from monocytes, fetal thymus, naive $CD4^+$ T cells, total $CD4^+$ T cells, activated total $CD4^+$ T cells, non-activated total $CD4^+$ T cells, naive $CD8^+$ T cells, total $CD8^+$ T cells, naive B cells and total B cells (Javierre et al. 2016) were used to identify genes interacting with index variants (cGenes). The data used to define functional genes (fGenes, pGenes and dGenes) were those used in the initial publication (Fang et al. 2019). Protein-protein interaction networks were determined according to the STRING database (Szklarczyk et al. 2017), requiring a confidence score ≥ 700.

4.2.8 Analytical tools and code

Unless otherwise noted, all statistical analyses and data visualization were performed using R version 3.6 (R Core Team 2017). All statistical tests based on symmetrically distributed test statistics were two-sided. No repeated measures data were analyzed in this study. All genotyped and ATAC-seq samples analyzed in the association tests represent distinct individuals. The R packages ggplot2 (Wickham 2016), cowplot (Wilke 2020), ggbio (Yin, Cook, and Lawrence 2012), GenomicRanges (Lawrence et al. 2013), gridExtra (Auguie 2017), RColorBrewer (Neuwirth 2014), and rtracklayer (Lawrence, Gentleman, and Carey 2009) were used for data visualization. Code used to generate the results presented in this chapter is available at `https://github.com/ccrobertson/t1d-immunochip-2020`.

4.3 Results

4.3.1 T1D credible variants are over-represented in accessible chromatin in T and B cells

ATAC-seq offers a high-resolution map of accessible chromatin with potential regulatory function (Buenrostro et al. 2015). Using publicly available (Calderon et al. 2019;

Ramos-rodríguez et al. 2019; Varshney et al. 2017; Jonsson et al. 2016) and newly generated ATAC-seq data from healthy donors, we assessed enrichment of 2,431 T1D credible variants (group posterior probability > 0.8) in accessible chromatin across diverse immune and non-immune cell types (including 25 primary immune cell types, pancreatic islets, and, as control cell types unlikely to be central to T1D etiology, fetal and adult cardiac fibroblasts). T1D credible variants were enriched in open chromatin in multiple primary immune cell types based on two complementary enrichment analysis approaches (Figure refmegasupfig9), with strong enrichment observed in stimulated $CD4^+$ effector T cells (Figure 4.6b). There was no enrichment in pancreatic islets ($p = 0.14$), the primary target of autoimmunity in T1D, even after exposure to proinflammatory cytokines ($p = 0.05$) or in cardiac fibroblasts ($p > 0.60$) (Figure 4.6).

Since many ATAC-seq peaks are present in both stimulated and unstimulated conditions, the enrichment scores are similar across conditions and may be underpowered to distinguish enrichment that is specific to a stimulated or unstimulated state. Therefore, for cell types with data available from unstimulated and stimulated conditions, we defined a subset of peaks that were significantly differentially accessible between the conditions (condition-specific peaks). Specifically, for each of 24 cell types, we defined peaks with significantly increased accessibility (FDR < 0.01) after stimulation ("stimulation-specific peaks") and peaks with significantly decreased accessibility after stimulation ("unstimulated-specific peaks"). Of 138,596 regions in the consensus peak set, Th17 cells had the highest proportion of stimulation-specific peaks (15.3%), while effector-memory $CD8^+$ T cells had the highest proportion of unstimulated-specific peaks (9.8%) (Supplementary Table 14). T1D credible variants were enriched in these condition-specific peaks in numerous cell types, with the largest enrichment in stimulation-specific peaks from effector $CD4^+$ T cells stimulated for 24 hours with anti-CD3/CD28 and human IL-2 (Figure 4.7). These results

indicate that T1D credible variants may contribute to islet autoimmunity, in part, by altering responses to T cell receptor signaling, co-stimulation, or cytokine signaling.

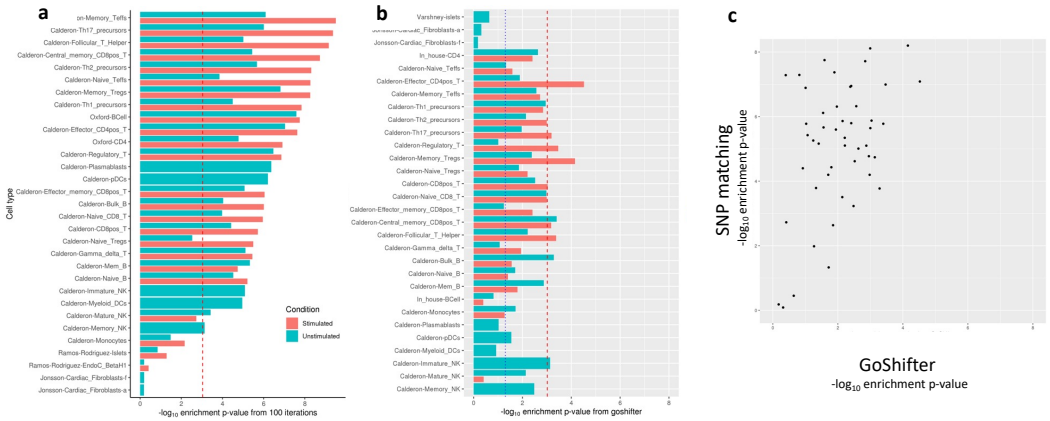

Figure 4.6: Enrichment of T1D credible variants in ATAC-seq peaks in each cell type (red bars, stimulated; green bars, unstimulated), red dashed line represents the Bonferroni significance threshold at the 5% level ($n = 2,431$ credible variants). (a) Enrichment analysis based on SNP-matching; (b) Enrichment analysis based on GoShifter; (c) Comparison of enrichments based on SNP-matching and GoShifter.

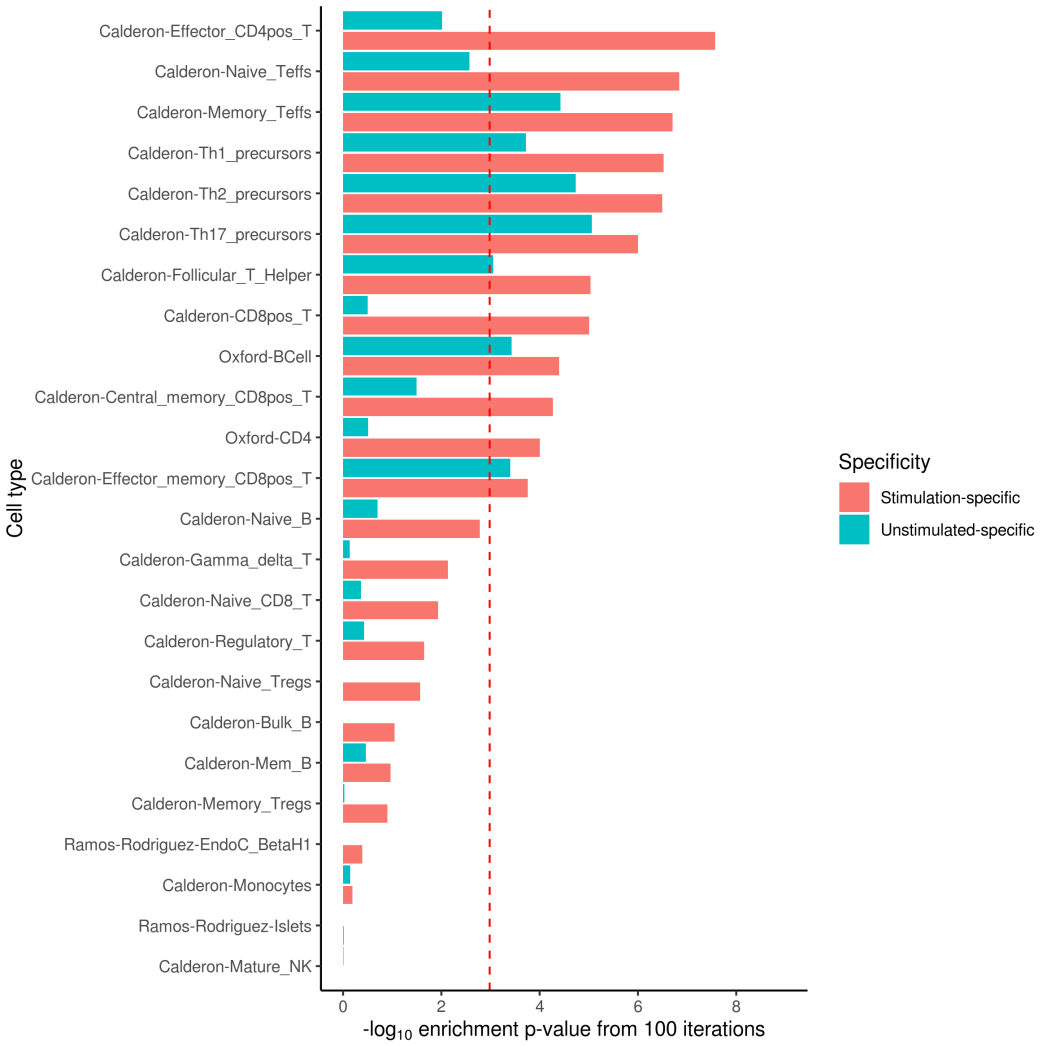

Figure 4.7: Enrichment of T1D credible variants in differentially open ATAC-seq peaks between stimulation conditions, defined from a consensus list of peaks. Red bars show differentially open peaks in stimulated cells; green bars show differentially open peaks in unstimulated cells. Red dashed line is the Bonferroni significance threshold at the 5% level ($n = 2,431$ credible variants).

4.3.2 Co-localization of T1D association with QTLs in immune cells

Chromatin accessibility profiles were generated across 115 participants ($n_{EUR} = 48$, $n_{AFR} = 67$) in primary CD4$^+$ T cells, the cell type in which accessible chromatin is most strongly enriched for T1D credible variants (Figures 4.6 and 4.7). We examined

additive effects of genotype on local chromatin accessibility (cis window < 1 Mb), identifying 11 "peaks" of chromatin accessibility significantly ($p < 5 \times 10^{-5}$) associated with T1D credible variants. Colocalization analysis of T1D association and caQTLs (R package coloc, Giambartolomei et al. 2014) identified five regions supporting a common causal variant underlying association with T1D and chromatin accessibility (PP.H4.abf > 0.8; Table 2). In all five regions, at least one T1D credible variant overlapped the caQTL-associated peak. Six of these "within-peak" credible variants were directly genotyped on the Immunochip, allowing us to examine allele-specific accessibility in heterozygous participants. At all six variants, the proportion of ATAC-seq reads from heterozygotes containing the alternative allele was consistent with the direction of the caQTL effect (Supplementary Table 15). When integrated with whole blood cis-eQTLs (Võsa et al. 2018), colocalization identified T1D candidate genes in four of five T1D-caQTL regions (PP.H4.abf > 0.8; Figure 4.8).

T1D lead variant[a]	β_{T1D}[b]	Peak	T1D credible variants in peak	caQTL lead variant[a]	β_{caQTL}[b]	P_{caQTL}	PP	Whole-blood *cis*-eQTLs[c]
rs71624119 (chr5:56144903:G:A)	−0.099	chr5:56147972-56149111	rs7731626	rs7731626 (chr5:56148856:G>A)	−0.5	2.4×10^{-9}	0.97	*ANKRD55* ($z=-58$; PP = 0.98) *IL6ST* ($z=-10$; PP = 0.98)
rs72928038 (chr6:90267049:G:A)	0.172	chr6:90266766-90267747	rs72928038	rs72928038 (chr6:90267049:G>A)	−1.0	3.9×10^{-16}	1.00	*BACH2* ($z=-21$; PP = 1)
rs2027299 (chr6:126364681:G:C)	0.147	chr6:126339725-126340580	rs9388486	rs1361262 (chr6:126380821:T>C)	−0.4	2.0×10^{-16}	0.87	*CENPW* ($z=-9.8$; PP = 0.82)
rs61555617[d] (chr12:56047884:TA:T)	0.257	chr12:56041256-56042638	rs705704 rs705705	rs705704 (chr12:56041628:G>A)	−0.2	1.1×10^{-15}	0.97	*GDF11* ($z=-7.5$[e]; PP = 0.97)
rs4900384 (chr14:98032614:A:G)	0.118	chr14:98018322-98019163	rs11628807 rs4383076 rs11628876 rs11160429	rs11628807 (chr14:98018774:T>G)	0.7	1.8×10^{-21}	0.95	-

Figure 4.8: T1D associations co-localizing with caQTLs in $CD4^+$ T cells. Five regions show co-localization between T1D and a caQTL with a co-localization posterior probability > 0.8. In all of these regions, at least one T1D credible variant overlaps the caQTL peak itself. In four regions, the T1D association also co-localizes with an eQTL for expression of one or more genes in whole blood. a, The T1D lead variant is the most-associated variant in the credible set, as defined by fine mapping (Supplementary Table 11); the caQTL lead variant is the most-associated variant with chromatin accessibility at the peak of interest. b, β_{T1D} refers to the effect size for the alternative allele of the T1D lead variant and β_{caQTL} refers to the effect size for the alternative allele of the caQTL lead variant. c, Whole-blood cis-eQTL statistics from eQTLGen for the T1D lead variant and co-localization with the T1D association. d, rs61555617 is referred to as rs796916887 in the Supplementary tables. e, The cis-eQTL statistics for rs61555617 are missing in eQTLGen; the reported *GDF11* cis-eQTL z-score is for the highly correlated variant rs705704. PP, posterior probability of co-localization between the QTL (eQTL or caQTL) and the T1D association (referred to in coloc documentation as "PP.H4.abf").

4.3.3 Functional annotation of T1D-associated variants in the *BACH2* region

Fine mapping of the *BACH2* locus refined the T1D association to two intronic variants, rs72928038:G>A and rs6908626:G>T (Figure 4.9a). The EUR minor alleles of rs72928038:G>A and rs6908626:G>T are associated with increased T1D risk (OR = 1.18; $p < 1\times10^{-20}$, $\text{MAF}_{EUR} = 0.18$). Chromatin-state annotations

across cell types from the BLUEPRINT Consortium and NIH Roadmap Epigenomics Project annotate rs72928038:G>A as overlapping a T cell-specific active enhancer and rs6908626:G>T as lying in the ubiquitous *BACH2* promoter (Figure 4.9b). Promoter-capture Hi-C data from diverse immune cell types (Javierre et al. 2016) indicates that the enhancer region containing rs72928038:G>A contacts the *BACH2* promoter in T cells (Figure 4.9c). Although weak interactions were observed in multiple T cell subtypes, only naive $CD4^+$ T cells had a significant interaction score.

In caQTL analysis, rs72928038:G>A is associated with decreased accessibility of the enhancer it overlaps (chr6:90266766-90267715) (Figure 4.9d, left), while rs6908626:G>T does not affect accessibility at the *BACH2* promoter (chr6:90294665-90297341) (Figure 4.9d, right). Similarly, among 14 subjects heterozygous for rs72928038:G>A, only 4% (5/121) of ATAC-seq reads overlapping that site contain the T1D risk allele (A) (Figure 4.9e, left, and Supplementary Table 15), suggesting it leads to restricted accessibility. In contrast, chromatin accessibility at rs6908626:G>T does not exhibit allelic bias in heterozygotes (Figure 4.9e, right). These data help to prioritize rs72928038:G>A, rather than rs6908626:G>T, as functionally relevant in $CD4^+$ T cells.

In eQTL studies, rs72928038:G>A is associated with decreased expression of *BACH2* in whole blood (Võsa et al. 2018) and purified immune cell types (Schmiedel et al. 2018). In the DICE consortium (Schmiedel et al. 2018), rs72928038:G>A is associated with decreased expression of *BACH2* in multiple cell types, with the strongest effects in naive $CD4^+$ and $CD8^+$ T cells. This result is consistent with the observation that the enhancer region overlapping rs72928038:G>A is accessible specifically in unstimulated bulk $CD4^+$, unstimulated bulk $CD8^+$, and naive $CD4^+$ T effector cells (Figure 4.9f). Both the enhancer caQTL and *BACH2* eQTL colocalize with T1D association (Figure 4.9g and Table 4.8).

The *BACH2* rs72928038:G>A variant overlaps binding sites for STAT1 and the

ETS family of transcription factors, based on canonical transcription factor binding motifs (Boyle et al. 2012). We performed super-shift electrophoretic mobility shift assay (EMSA) experiments of the DNA sequence flanking rs72928038:G>A that demonstrated allele-specific ETS1 binding, but no STAT1 binding (Figure 4.10). This result builds on experiments demonstrating allele-specific nuclear protein binding of rs72928038:G>A in Jurkat cells (Westra et al. 2018) . These data prioritize rs72928038:G>A as a likely functional variant in T cells and provide preliminary support for a candidate regulatory mechanism underlying the 6q15 region association with T1D. Specifically, we hypothesize that the rs72928038:G>A minor allele (A) disrupts ETS1 binding, which leads to decreased enhancer activity and *BACH2* expression in naive $CD4^+$ T cells.

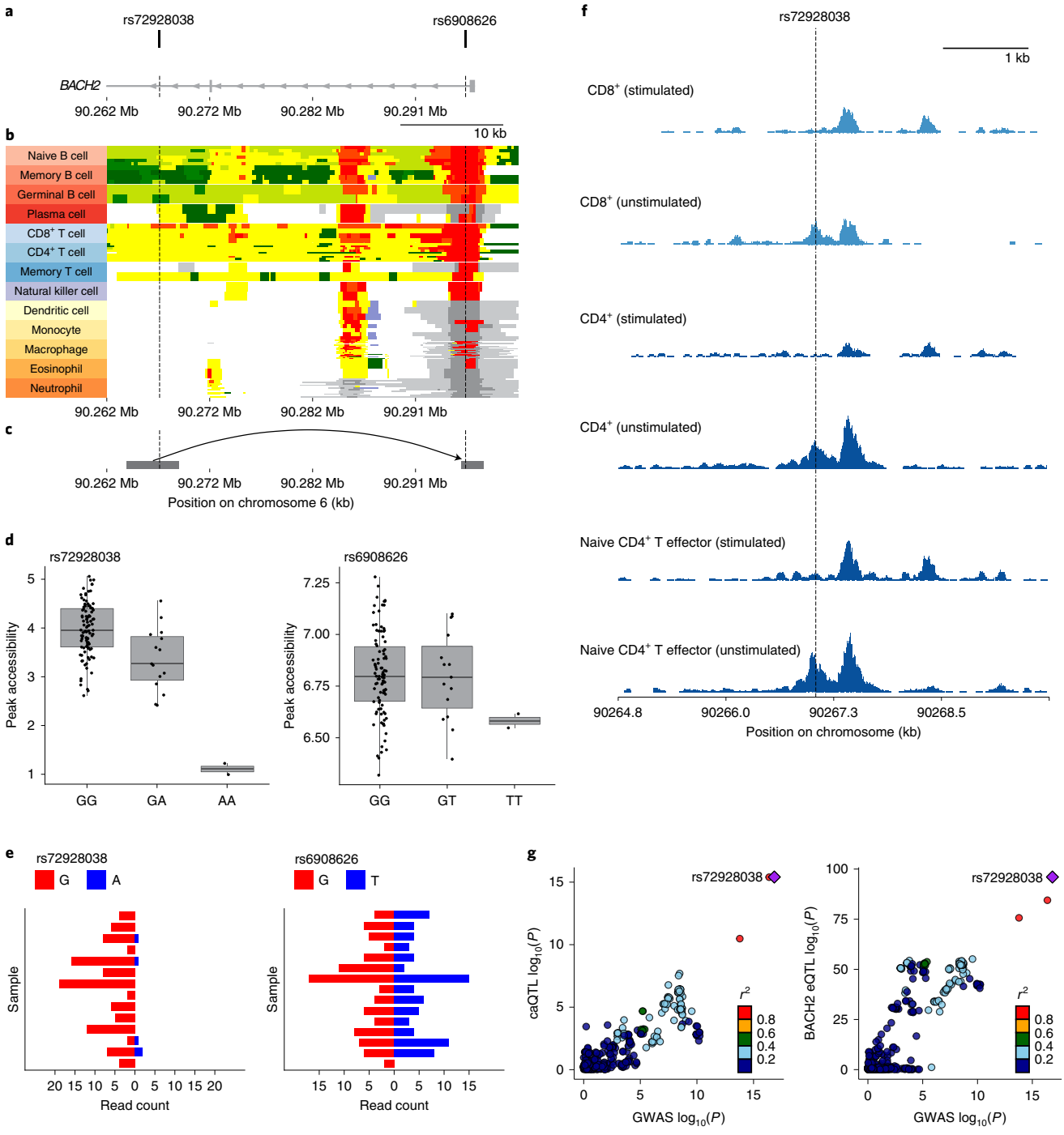
a
rs72928038
rs6908626
BACH2
90.262 Mb
90.272 Mb
90.282 Mb
90.291 Mb
b
10 kb
Naive B cell
Memory B cell
Germinal B cell
Plasma cell
CD8+ T cell
CD4+ T cell
Memory T cell
Natural killer cell
Dendritic cell
Monocyte
Macrophage
Eosinophil
Neutrophil
c
Position on chromosome 6 (kb)
d
rs72928038
rs6908626
Peak accessibility
GG
GA
AA
GT
TT
e
G
A
T
Sample
Read count
f
1 kb
CD8+ (stimulated)
CD8+ (unstimulated)
CD4+ (stimulated)
CD4+ (unstimulated)
Naive CD4+ T effector (stimulated)
Naive CD4+ T effector (unstimulated)
90264.8
90266.0
90267.3
90268.5
Position on chromosome (kb)
g
caQTL log10(P)
BACH2 eQTL log10(P)
GWAS log10(P)
r2
0.8
0.6
0.4
0.2

Figure 4.9: Functional annotation of T1D-associated variants in the *BACH2* region. a - c, Position of T1D credible variants (rs72928038:G>A and rs6908626:G>T) relative to (a) the introns and exons of *BACH2*; (b) chromHMM tracks across diverse immune-cell types from the BLUEPRINT consortium (red, active promoter; orange, distal active promoter; dark green, transcription; light green, genic enhancer; yellow, enhancer; white, quiescent; light gray, Polycomb repressed; dark gray, repressed; and blue, heterochromatin); and (c) interactions with the *BACH2* promoter in published promoter-capture Hi-C data from naive $CD4^+$ T cells (the gray squares indicate boundaries of target (left) and bait (right)). The chromatin coordinates and the scale are identical and aligned. d, Accessibility of regions overlapping rs72928038:G>A (left) and rs6908626:G>T (right) according to genotype. Peak accessibility is quantified as the normalized transposase cut frequency; center line, median; box limits, upper and lower quartiles; whiskers, 1.5x the interquartile range ($n = 115$ individuals). e, Allele-specific accessibility of chromatin within heterozygous individuals at rs72928038:G>A (left; $n = 14$ heterozygous individuals) and rs6908626:G>T (right; $n = 15$ heterozygous individuals). f, Chromatin-accessibility profiles in the region overlapping rs72928038:G>A across resting and activated $CD4^+$ and $CD8^+$ T cells (published data). The height of the tracks represents the transposase cut frequency; all tracks are plotted using the same vertical scale. g, LocusCompare plots showing co-localization between T1D association, the caQTL for chr6:90266766-90267715 (left) and the eQTL for *BACH2* (right).

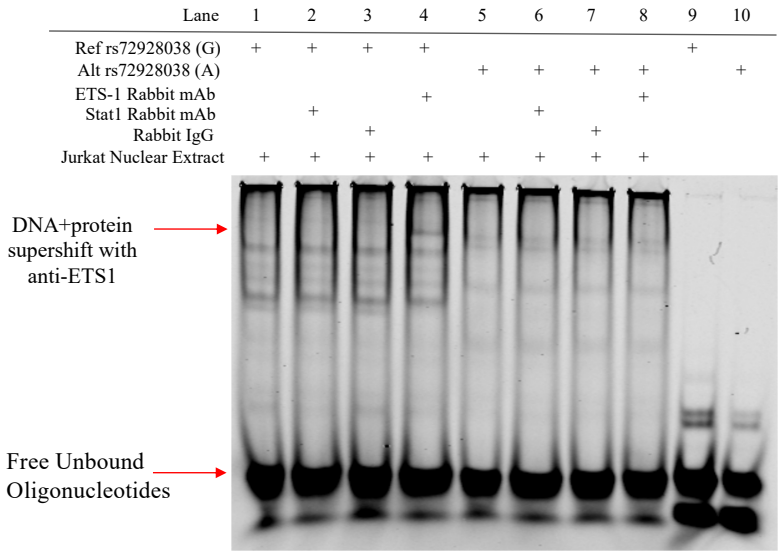

Figure 4.10: rs72928038 with ETS-1 antibody supershift Electrophoretic Mobility Shift Assay (EMSA). Lane 1-4 and 9 contains the reference allele (G) of rs72928038 labeled probe. Lane 5-8 and 10 contains the alternative allele (A) of rs72928038 labeled probe. Rabbit IgG was added to lane 3 and 7 as negative controls for the supershift assay. Lane 9 and 10 are negative controls. The ETS-1 supershift EMSA demonstrates an allele-specific supershift with rs72928038 G allele, while a shift is not observed with the A allele probe. Specifically, in lane 4, we see the appearance of a band that is not present in lanes to 1 and 3, which suggests ETS1 binding of the labeled probe containing the rs72928038 G allele. Meanwhile, we do not see any differences in band patterns between lanes 8 and lanes 5 and 7, which suggests that there is no ETS1 binding of the labeled probe containing the rs72928038 A allele. Likewise, we do not see any new bands in lane 2 relative to lanes 1 and 3, or lane 6 relative to lanes 5 and 7, which suggests that STAT1 does not bind the labeled probe for either rs72928038 allele. Experiments showing allele-specific ETS1 binding were repeated 4 times. Experiments showing lack of STAT1 binding were repeated three times.

4.3.4 T1D drug target identification

To identify potential T1D therapeutic targets with human genetic support, we used the Priority Index (Pi) algorithm (Fang et al. 2019), which integrates genetic association results with genome annotations, regulatory maps, and protein-protein networks. Using improved T1D association statistics and additional eQTL resources from whole blood (Võsa et al. 2018), we identified 50 highly-ranked gene targets (Supplementary

Table 16). These targets include 26 "seed genes" (implicated by T1D-associated loci through proximity, eQTL effects, or chromatin looping) and 24 non-seed genes (not in T1D regions but highly connected to T1D seed genes in immune protein networks). Although we excluded variants in the MHC region from algorithm input, the networks implicated by non-MHC seed genes led to prioritization of *HLA-DRB1*, an established T1D risk factor. Among the top 50 gene targets, 13 were not previously implicated by Pi analyses (*STAT4*, *RGS1*, *CXCR6*, *IL23A*, *PTPN22*, *NFKB1*, *MAPK3*, *EPOR*, *DGKQ*, *GALT*, *IL12RB1*, *IL12RB2*, *IL6R*), and 12 have been targeted in clinical trials for autoimmune diseases (*IL2RA*, *IL6ST*, *IL6R*, *TYK2*, *IFNAR2*, *JAK2*, *IL12B*, *IL23A*, *IL2RG*, *JAK3*, *JAK1* and *IL2RB*). T1D susceptibility alleles may alter expression of gene targets in either direction, and gene regulatory effects may be seen across multiple major immune cell populations or be restricted to a single cell type (Figure 4.11). For example, T1D risk alleles are associated with increased expression of *MAPK3* and *DGKQ* but decreased expression of *TYK2* across multiple major immune cell populations. In contrast, risk alleles decrease expression of *RGS1* across most immune cell types but increase expression specifically in $CD8^+$ T cells. The directionality and cell type-specificity of gene regulatory effects associated with T1D risk alleles may inform therapeutic target considerations.

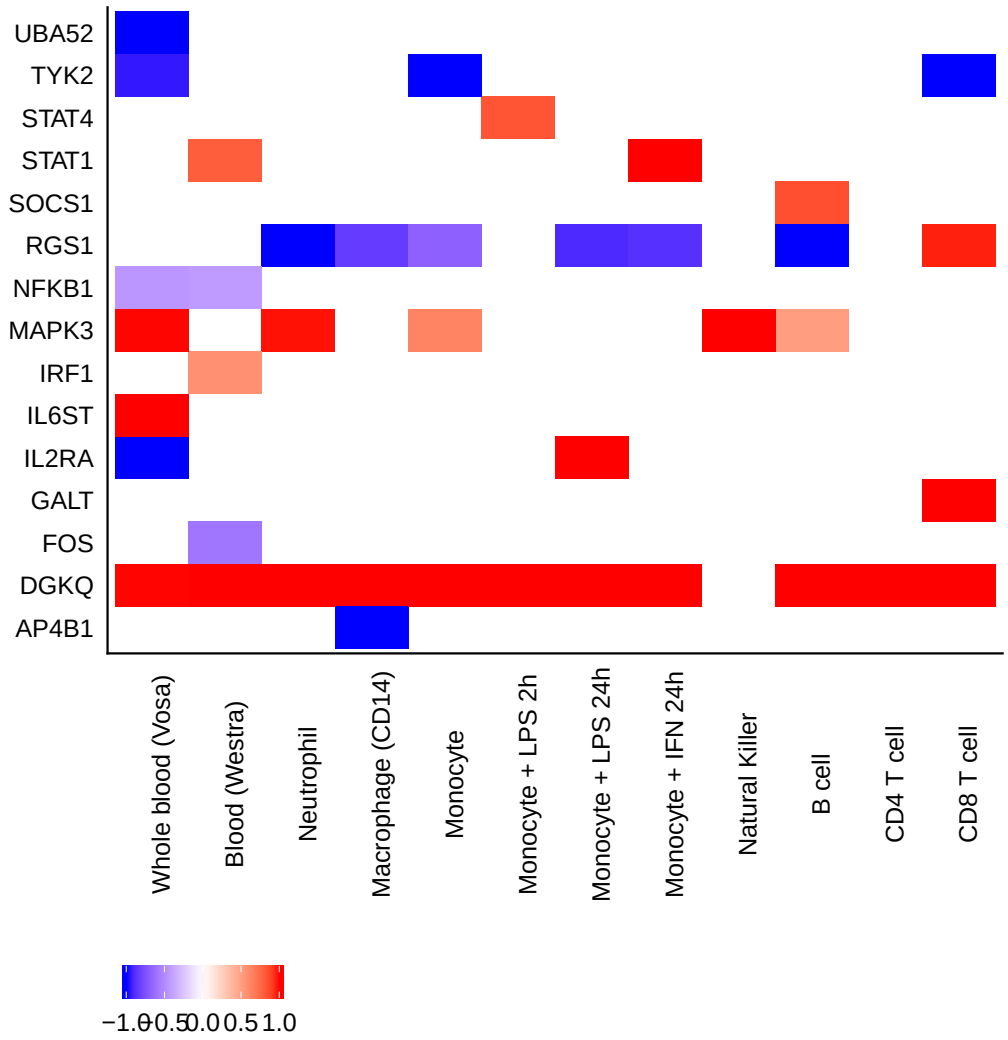

Figure 4.11: Direction of eQTL effects among Priority Index target genes across immune cell populations. For each of 16 target genes where eQTLs colocalize with T1D association, the direction of effect and strength of co-localization is shown for each of 12 cell contexts. Red and blue squares indicate the T1D risk allele is associated with increased or decreased gene expression, respectively. Intensity of the color reflects the posterior probability (PP_{ABF}) for co-localization between the eQTL and T1D association, such that darker rectangles imply stronger evidence of a shared causal variant.

4.4 Discussion

We assessed the intersection of T1D-associated variants with regions of putative regulatory function with public and newly generated ATAC-seq data from diverse

cell types and states, demonstrating that T1D credible variants were enriched in stimulation-responsive open chromatin peaks in $CD4^+$ T cells. We assessed co-localization of T1D associations with $CD4^+$ T cell caQTLs to generate mechanistic hypotheses centered on this highly relevant cell type. Finally, we identified potential T1D drug targets for use in prevention trials. Experimental follow-up studies are required to test these hypotheses and further dissect the mechanisms altering T1D risk in each region.

Despite the enrichment of credible variants in $CD4^+$ T-cell open chromatin, only five of 52 fine-mapped T1D associations could be explained by a co-localized caQTL. This result is consistent with work exploring the functional effects of variants associated with immune traits (Chun et al. 2017). One explanation is limited power in QTL discovery due to small sample sizes or imprecise cell types (Chun et al. 2017; Hukku et al. 2020). The analysis of more refined cell types using single-cell approaches, for both enrichment analyses and QTL discovery, may lead to additional discoveries (Chiou et al. 2021; Paola Benaglio, Jacklyn Newsome, Jee Yun Han, Joshua Chiou, Anthony Aylward, Sierra Corban, Mei-Lin Okino, Jaspreet Kaur, David U Gorkin 2020). Nevertheless, although this approach may lack sensitivity, the five regions showing co-localization between caQTL and T1D associations prioritize variants with regulatory effects that represent realistic targets for experimental follow-up. In particular, within-peak credible variants with consistent caQTL effects and allele-specific accessibility, although not definitively causal, provide high-priority candidate variants for functional follow-up. As four of the five T1D associations that co-localize with caQTLs also co-localize with whole-blood eQTLs, these regions offer hypotheses for how causal variants influence disease risk through their effects on regulatory element activity and gene expression in T1D-relevent cell types. Credible variants were restricted to Immunochip content, which may explain the absence of T1D-variant enrichment in the open chromatin of non-immune-cell types (for example, pancreatic

islets) (Aylward et al. 2018; Dooley et al. 2016).

In the 5q11.2 region, fine mapping and caQTL co-localization point to the within-peak variant rs7731626 (G>A) as a potential causal variant for T1D. This result complements a recent regulatory QTL fine-mapping study that highlighted the same variant as likely to be functional in T cells (Kundu et al. 2020). In addition, the T1D association co-localizes with eQTLs for both *ANKRD55* and *IL6ST*, mirroring results in multiple sclerosis, Crohn's disease and rheumatoid arthritis (Chun et al. 2017). The region overlapping rs7731626 (G>A) loops to the *IL6ST* promoter in $CD4^+$ T cells, according to promoter-capture Hi-C data (Javierre et al. 2016). Although we did not find evidence that rs7731626 (G>A) loops to the canonical transcription start site for *ANKRD55*, nascent RNA-sequencing data suggest it overlaps the 5' end of the transcriptionally active region of *ANKRD55* in human T cells (Danko et al. 2018), consistent with a potential regulatory role.

We highlight the *BACH2* region on chromosome 6q15 as an example of unbiased QTL co-localization that leads to hypotheses for functional mechanisms driving variant-T1D association. We hypothesize that rs72928038 (G>A), the T1D-associated allele, abolishes ETS1 binding at an enhancer that promotes *BACH2* expression in naive $CD4^+$ T cells. *BACH2* encodes the transcription factor from the BTB-basic leucine zipper family, BACH2, which has established roles in B- and T-cell biology, including maintenance of the naive T-cell state (Tsukumo et al. 2013; Roychoudhuri et al. 2016). *BACH2* haploinsufficiency has been shown to cause congenital autoimmunity and immunodeficiency (Afzali et al. 2017), demonstrating that a functioning human immune system depends on *BACH2* expression in a dose-dependent manner. In addition to cis-effects on *BACH2* expression, rs72928038 (G>A) is associated with altered expression of 39 distal genes in whole blood (Võsa et al. 2018), including seven genes in autoimmune disease-associated regions. These observations raise the hypothesis that the minor A allele at rs72928038 (G>A) increases T1D risk

by reducing *BACH2* expression in a precise cellular context (for example, the naive T-cell state). This effect may lead to shifts in BACH2-regulated transcriptional programs, thereby altering T cell lineage differentiation in response to antigen exposure.

Previous studies demonstrated shared genetic risk across autoimmune diseases (Onengut-Gumuscu et al. 2015; Cotsapas et al. 2011) and suggest the potential for repurposing drugs to treat or prevent T1D. Our priority-index analysis identified 12 targets that have been the focus of clinical trials for the treatment of autoimmune diseases. One example is *IL23A*, which has been successfully targeted in the treatment of inflammatory bowel disease (Faegan et al. 2018) and psoriasis (Fotiadou et al. 2018). The IL-23 inhibitors are being explored for use in T1D (ClinicalTrials.gov identifiers NCT02204397 and NCT03941132). Our results provide genetic support for these trials. Similarly, *JAK1*, *JAK2* and *JAK3* were implicated in T1D etiology in our analysis. JAK inhibitors are safe and effective in the treatment of rheumatoid arthritis (Wollenhaupt et al. 2019) and ulcerative colitis (Sandborn et al. 2017).

Finally, this study presents the first well-powered convincing genetic evidence linking interleukin-6 (IL-6), a cytokine with known roles in multiple autoimmune diseases, to T1D etiology. The IL-6R complex consists of two essential subunits: the alpha subunit (encoded by *IL6R*) and the signal-transducing subunit (encoded by *IL6ST*). Both the *IL6ST* and *IL6R* regions were identified here as T1D-associated at genome-wide significance (Figure 3.13), and both *IL6ST* and *IL6R* were prioritized by the Pi analysis. *IL6ST* is implicated by QTL co-localization and the lead T1D variant near *IL6R* (rs2229238 (T>C)) is an eQTL for *IL6R* expression in whole blood (formal co-localization was not assessed as the *IL6R* region is not densely covered by the Immunochip). Based on the current evidence, we cannot say that *IL6ST* and *IL6R* are T1D causal genes. The associations in each region may be unrelated and due to different causal genes - for example, the association near *IL6ST* also co-localizes with an eQTL for *ANKRD55*. However, we note that the humanized IL-6R antagonist

monoclonal antibody tocilizumab is an approved treatment for rheumatoid arthritis and systemic juvenile idiopathic arthritis, both of which share substantial genetic effects with T1D (3.14, Onengut-Gumuscu et al. 2015), and a trial of this drug in recently diagnosed T1D cases is underway (ClinicalTrials.gov identifier NCT02293837). Surprisingly, we showed that the lead T1D variant near *IL6R* (rs2229238 (T>C)) tags a causal variant distinct from the nonsynonymous variant in *IL6R*, rs2228145 (A>C) (p.Asp358Ala), that is thought to drive the association in rheumatoid arthritis (Okada et al. 2014), suggesting potentially different mechanisms altering disease risk in this region. The recent success of anti-CD3 therapy, after 40 years of study through experimental models and clinical trials targeting different patient subgroups and time points relative to disease diagnosis (Gaglia and Kissler 2019), highlights both the challenges and hopes for translating target identification to efficacious clinical outcomes in T1D.

Chapter 5

Future directions

In this thesis, I have explored the genetic basis of T1D through multiple collaborative projects, with my contributions highlighted in each Chapter. In Chapter 2, I evaluated HLA association with T1D in an African American cohort, identifying shared and unique susceptibility alleles between African- and European-ancestry populations. In Chapter 3, we discovered 36 new regions associated with T1D at genome-wide significance and defined a total of 152 regions associated with T1D using a false discovery rate (FDR) approach. Through this work, we have begun to clarify the genetic factors contributing to T1D risk beyond the canonical pediatric onset Northern European population, which has been the focus of the vast majority of T1D genetic studies to date. Finally, in Chapters 3 and 4, we defined likely causal variants in 52 T1D-associated regions and reveal potential gene regulatory mechanisms in five of these regions. Here, I will discuss how this work may inform future efforts in T1D genetics and its role in precision medicine.

5.1 The future of genetic discovery in T1D

The fine-mapping ImmunoChip array had been used previously in a large-scale T1D study in European-ancestry samples (Onengut-Gumuscu et al. 2015). We found 36

new T1D loci due increased sample size, inclusion of genetically diverse subjects, and imputation. In most of the novel loci, lead variants were common alleles with modest effects on disease risk. However, fine-mapping revealed extensive allelic heterogeneity, and at some loci with multiple causal variants, secondary associations map to low-frequency large effect variants. For example, we identified protective low-frequency nonsynonymous and splice variants in *IFIH1* and *TYK2* (MAF ranging from 0.5 to 5%, OR about 0.6).

Meanwhile, in African-ancestry subjects, we identified a low-frequency variant with very large effect on T1D risk near the gene *MTF2* (MAF = 1%, OR = 2.9). A recent genome-wide meta-analysis of existing European T1D cohorts with two population biobanks, from the United Kingdom (Bycroft et al. 2018) and Finland (`www.finngen.fi`), identified 15 additional loci (Chiou et al. 2021). In this genome-wide meta-analysis (Chiou et al. 2021), discovery exhibited similar trends to our results (Figure 5.1), where all but one novel region had a modest effect size (OR < 1.5), but fine-mapping revealed additional low-frequency variants with moderate or large effects in known T1D regions. Common themes from our work and others' can help frame expectations for future genetics studies of T1D.

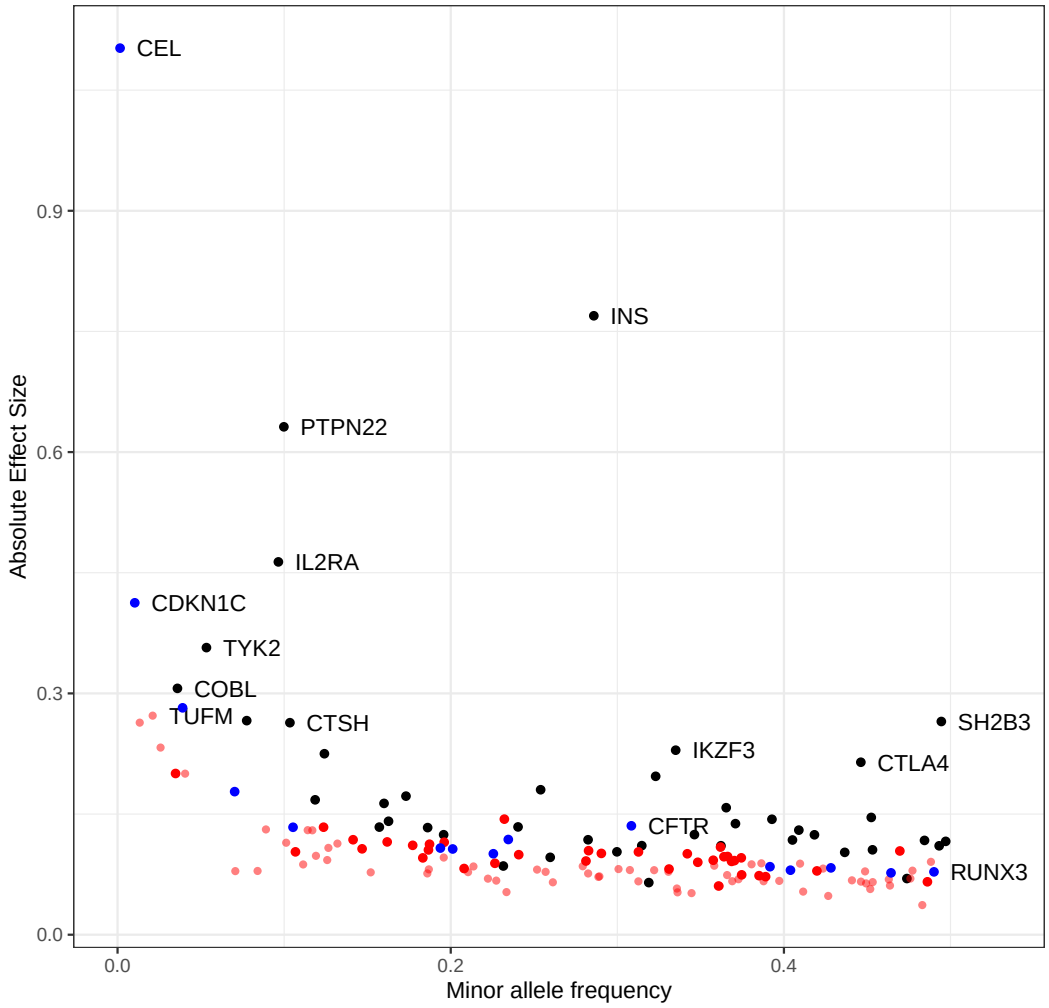

Figure 5.1: European-ancestry allele frequency and effect size distribution for variants associated with type 1 diabetes (T1D). Each point represents a lead variant from one T1D-associated region. Variants are positioned according to their absolute effect size (y-axis = log(OR)) and minor allele frequency (x-axis) in European case-control analyses. red, associated with T1D for the first time in our meta-analysis; blue, associated with T1D for the first time in Chiou et al. 2021; black, previously associated with T1D.

In aggregate, these studies suggest that analysis of larger European cohorts will primarily identify common variants with increasingly smaller effects on T1D risk. This trend will likely persist even if European analysis is expanded to genome-wide analysis, instead of fine-mapping approaches. Regardless of genotyping platform, we are unlikely to identify additional common variants with moderate or large effects

on T1D risk in European populations. However, our work suggests that we are only beginning to scratch the surface of genetic contributions to T1D in African-ancestry populations (Figure 3.22), which harbor more genetic diversity than any other human ancestral populations (1000 Genomes Project Consortium 2015). The genetic architecture for T1D in other groups, including Hispanic- and Asian-ancestry populations, are even less defined. Due to limited sample sizes, the effect size estimates for the non-European groups in our study are imprecise (wide confidence intervals). Moreover, our analyses were restricted to regions previously implicated in autoimmune disease studies of largely European-ancestry cohorts (i.e., the ImmunoChip). Thus, there may be many additional variants common in African-, Hispanic-, and Asian-ancestry groups that contribute to T1D risk but have escaped identification due to their low frequency or absence in Europeans.

Since the publication of our African-ancestry analyses (Onengut-Gumuscu et al. 2019), several published reports have highlighted the challenge of genetic risk prediction portability across populations (Martin et al. 2019; Privé et al. 2021; Lam et al. 2019), and diverse approaches to address this issue have been proposed (Amariuta et al. 2020; Wand et al. 2021). Ultimately, sufficient representation of diverse groups in genetic studies will be required to develop generalizable genetic prediction models, and to ensure equitable application of genetic medicine across all patients (Wojcik et al. 2019b). Currently, large-scale efforts are underway to enrich for underrepresented groups in large population-based studies (e.g., All of Us (All of Us Research Program Investigators 2019), and Million Veteran's Program (Gaziano et al. 2016)). However, due to the low prevalence of T1D, population-based studies are typically under-powered. Even the UK Biobank (Bycroft et al. 2018), which recruited about 500,000 subjects from the United Kingdom, contains fewer than 1,500 T1D cases (Chiou et al. 2021; Thomas et al. 2018). Thus, accurate prediction of T1D across the full spectrum of patient groups will require additional targeted efforts to generate

cohorts of non-European T1D cases.

Larger studies will also identify rare variants with large effects on T1D. In particular, our work suggests substantial allelic heterogeneity at known T1D loci, frequently with multiple rare variants modifying a common disease-proximal gene. However, as our haplotype and fine-mapping analyses demonstrate, deciphering the causal variants in these regions is far from straightforward. This challenge is illustrated by the chromosome region 2q24.2, which contains the established T1D gene *IFIH1*. Both our analysis and and those of others (e.g., Chiou et al. 2021) identified allelic heterogeneity in the *IFIH1* region, and both studies conclude that there are likely multiple rare or low-frequency protein-altering variants in *IFIH1* that each provide substantial protection from T1D. However, the specific variants prioritized in our studies differed. In particular, a nonsynonymous variant in *IFIH1* that we had excluded from our analysis due to its low frequency (rs75671397, p.Asn160Asp, MAF = 0.002, OR = 0.35) was prioritized by another group (Chiou et al. 2021). Due to the complexity of association in this region, and the imperfect nature of imputation, particularly of lower frequency variants, it remains unclear which variants are causal. In regions like this, with complex genetic architecture, direct observation (i.e., sequencing) of inherited haplotypes may provide increased confidence in causal variant prioritization.

Together, these conclusions suggest that the largest gains in biological insight for T1D will most likely come from analysis of groups previously underrepresented in T1D research and using sequencing-based platforms that can accurately measure low frequency and rare variation. Emphasis on these two goals may be synergistic in some contexts. For example, linkage disequilibrium decays more quickly in individuals of African ancestry (Figure 3.2), which both makes it more difficult to accurately infer rare and low-frequency variants using reference haplotypes (i.e., imputation is harder) and makes it easier to resolve causal variants underlying an association (i.e., fine-mapping is easier). Thus, whole genome sequencing analysis of large numbers of

T1D cases with African ancestry will facilitate both rare variant discovery and improve fine-mapping resolution. Combined with the fact that rare and low-frequency variants often have more potent effects on the causal gene through which they influence disease, this approach may provide an especially efficient path from genetic mapping to biological insight.

5.2 Mapping T1D associations to causal variants and genes

To prioritize potential regulatory mechanisms underlying non-coding T1D-associated variants, we integrated T1D credible sets with chromatin accessibility profiles and chromatin accessibility QTLs from $CD4^+$ T cells. While this approach provided hypotheses for molecular function of T1D variants in five regions, the majority of T1D-associated regions remain unexplained.

More generally, while functional genomics has been useful for interpreting results from genetic association studies (Cano-Gamez and Trynka 2020), only a small fraction of autoimmune disease-associated loci are explained by co-localization with molecular QTLs (Chun et al. 2017). Furthermore, experimental follow-up to confirm the proximal regulatory effect of a candidate causal variant (e.g., luciferase assays to evaluate allele-specific enhancer activity) are time-consuming and low throughput. Considering a typical credible set at a GWAS locus contains up to dozens of potentially causal variants, approaches to functionally test many non-coding genetic variants for regulatory effects simultaneously will be essential for increasing the efficiency of GWAS follow-up. In particular, massively parallel reporter assays (Melnikov et al. 2012; Arnold et al. 2013) and high throughput genome editing followed by single cell sequencing (Fulco et al. 2016; Schraivogel et al. 2020; Pan et al. 2020) will likely play a role (Morris et al. 2021). Integration of results from these perturbation experiments

with existing molecular QTL data sets may provide a powerful and efficient pipeline for systematically nominating causal genes and variants at GWAS loci.

5.3 Mechanisms for HLA association with T1D

This work supports common mechanisms underlying HLA-associated T1D risk across ancestral populations. While most of T1D risk in the HLA region is likely mediated by coding sequence changes in HLA class I and II molecules, additional mechanisms may be at play. Recent work suggests that other genes or non-coding regulatory variation could be underlying association with T1D. Variation in the *BTNL2* gene, located near the HLA class II genes, contributes to T1D risk even after controlling for HLA DR-DQ genotype (Hippich et al. 2019). Using high-depth RNA-sequencing, classical HLA alleles were found to be dynamically regulated, showing varying levels of allele-specific expression in stimulated T cells over time (Gutierrez-Arcelus et al. 2020). Meanwhile, single cell analysis of T1D pancreas tissue identified elevated MHC class II expression on exocrine ductal cells (Fasolino et al. 2021).

Our analyses were based on high-density SNP genotypes from the ImmunoChip (about 5,000 SNPs in the 6 Mb MHC region), as well as, imputed 2-field HLA alleles and HLA gene amino acid sequences. Higher resolution sequence-based HLA typing or improved imputation, using an expanded HLA imputation panel (Luo et al. 2020) or novel HLA imputation methods (Naito et al. 2021), may provide additional insights about the basis of T1D risk in the region. However, due to the complex linkage disequilibrium structure, causal variants and mechanisms are likely to remain unclear even if it becomes possible to obtain entire nucleotide-resolution HLA haplotypes from large numbers of T1D cases. Thus, creative approaches integrating diverse data sets (e.g., RNA and protein level expression of HLA alleles, peptide-MHC complexes, and TCR repertoires) from carefully selected patient-derived samples (e.g.,

antigen presenting cells from the draining pancreatic lymph nodes), may be required to elucidate the full spectrum of mechanisms underlying association between MHC genotypes and T1D risk in human populations.

5.4 Precision medicine for T1D

As approaches to T1D risk stratification improve, we will have the opportunity to prevent immune-mediated destruction of β cells in a large proportion of individuals who would otherwise go on to develop T1D. T1D risk prediction has already been improved by accounting for interaction effects and correlation structure of genetic variants in the MHC region (Sharp et al. 2019; Zhao et al. 2017) and combining genetic risk scores with other non-genetic factors associated with T1D (Ferrat et al. 2020). Genome-wide use of variation to predict genetic risk (i.e., Polygenic Risk Score (PRS) and its transferability across ancestry groups is an active area of research. Additional work to integrate available data to improve T1D prediction will increase the value of genetic risk models in clinical settings (Chung et al. 2020).

Etiologically, T1D is a complex disease, influenced by many genetic and environmental risk factors. However, the end result of this complex process is relatively straightforward: individuals with T1D can no longer regulate their blood glucose because they lack functional β cells. The simplicity of the problem is confirmed by the effectiveness of pancreas and islet transplantation. After pancreas transplantation, many patients remain insulin-independent for multiple years (Gruessner and Gruessner 2013). Allogeneic islet transplants, where islets are isolated from organ donors and infused into the recipient, have also been effective in restoring normal glucose control (Shapiro, Pokrywczynska, and Ricordi 2017). Unfortunately, in both cases, recipients must take steroids and immunosuppressants to prevent transplant rejection. Due to the risks associated with chronic immunosuppression (Rama and Grinyo

2010), patients are only eligible for pancreas or islet transplantation if they are already receiving another organ transplant (e.g., a kidney transplant due to diabetes-induced end stage renal disease) or if they have a dangerous condition known as "hypoglycemia unawareness" (Gruessner and Gruessner 2013; Shapiro, Pokrywczynska, and Ricordi 2017).

Currently, researchers are working on approaches to differentiate patient-derived stem cells into functional islets (Pagliuca et al. 2014; Yoshihara et al. 2020), which could then be used for autologous islet cell transplantation. If successful, this approach could restore glucose homeostasis and insulin independence without introducing foreign genetic material into the recipient, eliminating the need for chronic immunosuppression. However, the same autoimmune process that precedes T1D onset may recur and eventually destroy the transplanted β cells. Thus, a better understanding of the autoimmune mechanisms driving initiation of islet autoimmunity and progression to T1D are still important for effective treatment of individuals with T1D.

Numerous therapeutic interventions are able to prevent autoimmune diabetes in spontaneous animals models but have no effect on T1D progression in humans (Bowman, Leiter, and Atkinson 1994). This observation suggests a complexity or heterogeneity in T1D etiology in humans that is not recapitulated in existing animal models. Defining the human genetic factors contributing to T1D risk may provide opportunities to infer disease sub-types (Dahl and Zaitlen 2020) and tailor therapeutic interventions to the patients who will benefit from them.

Already, studies have begun to identify correlations between T1D-associated HLA types and molecularly- or clinically-distinct T1D subgroups (**krischer20156**; Inshaw et al. 2020). The full spectrum of genetic variation contributing to T1D may eventually be leveraged in a similar way. This opportunity has been framed in the context of coronary artery disease (Khera and Kathiresan 2017), motivated by the belief that, in most patients, complex disease is driven by a "quantitative blend of causal fac-

tors." Under this model, we may think of genetically complex diseases as a blend of ingredients, where each ingredient is a molecular pathway involved in the clinical phenotype. Since the genetic factors driving disease are dispersed widely across the genome, it can be difficult to delineate the common set of pathways they disrupt. Moreover, since most causal variants are in non-coding regions, assigning individual variants to genes and pathways remains a challenge. However, as molecular tools are developed to efficiently and systematically map disease-associated variants to causal genes, we may eventually use genome-wide profiles to estimate, not just the risk of disease onset, but also the relative contribution of relevant pathways to the disease process in individual patients.

With these goals in mind, one can envision a future where abrupt, unexpected T1D onset becomes rare due to population screening incorporating both genetic and non-genetic susceptibility factors, close monitoring of high-risk individuals, and targeted intervention upon early evidence of autoimmune activity. Meanwhile, in the rare individual who develops T1D despite a low-risk profile, autologous islet cell transplantation can replenish lost β cells and targeted immune intervention can prevent disease recurrence.

In conclusion, the future of T1D research may be focused on genetic discovery in populations previously underrepresented in T1D research, efficient mapping of genetic risk variants to causal genes and cell types, delineating major disease pathways and mechanisms disrupted by T1D-associated variants, and quantifying the relative contribution of such pathways to the disease process for individual patients. Implementation of this work will guide population screening, patient stratification and sub-typing, therapeutic target identification, and participant recruitment to immune intervention trials for T1D prevention.

Bibliography

1000 Genomes Project Consortium et al. (2015). "A global reference for human genetic variation". *Nature* 526.7571, p. 68.

Afzali, Behdad et al. (2017). "BACH2 immunodeficiency illustrates an association between super-enhancers and haploinsufficiency". *Nature Immunology* 18.7, pp. 813–823. DOI: 10.1038/ni.3753.

AH, I Barnett et al. (1981). "Diabetes in identical twins. A study of 200 pairs". *Diabetologia* 20, pp. 87–93.

All of Us Research Program Investigators (2019). "The "All of Us" research program". *New England Journal of Medicine* 381.7, pp. 668–676.

Allis, C David and Thomas Jenuwein (2016). "The molecular hallmarks of epigenetic control". *Nature Reviews Genetics* 17.8, pp. 487–500.

Amariuta, Tiffany et al. (2020). "Improving the trans-ancestry portability of polygenic risk scores by prioritizing variants in predicted cell-type-specific regulatory elements". *Nature Genetics* 52.12, pp. 1346–1354.

Amberger, Joanna S et al. (2015). "OMIM. org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders". *Nucleic Acids Research* 43.D1, pp. D789–D798.

Andiappan, Anand Kumar et al. (2015). "Genome-wide analysis of the genetic regulation of gene expression in human neutrophils". *Nature Communications* 6, p. 7971. DOI: 10.1038/ncomms8971.

Arif, Sefina et al. (2014). "Blood and islet phenotypes indicate immunological heterogeneity in type 1 diabetes". *Diabetes* 63.11, pp. 3835–3845.

Arnold, Cosmas D et al. (2013). "Genome-wide quantitative enhancer activity maps identified by STARR-seq". *Science* 339.6123, pp. 1074–1077.

Asimit, Jennifer L et al. (2019). "Stochastic search and joint fine-mapping increases accuracy and identifies previously unreported associations in immune-mediated diseases". *Nature Communications* 10.3216. ISSN: 2041-1723. DOI: 10.1038/s41467-019-11271-0.

Atkinson, Mark A and George S Eisenbarth (2001). "Type 1 diabetes: new perspectives on disease pathogenesis and treatment". *The Lancet* 358.9277, pp. 221–229.

Auguie, Baptiste (2017). *gridExtra: Miscellaneous Functions for "Grid" Graphics*. R package version 2.3. URL: https://CRAN.R-project.org/package=gridExtra.

Aylward, Anthony et al. (2018). "Shared genetic risk contributes to type 1 and type 2 diabetes etiology". *Human Molecular Genetics*. ISSN: 0964-6906. DOI: 10.1093/hmg/ddy314.

Babon, Jenny Aurielle B et al. (2016). "Analysis of self-antigen specificity of islet-infiltrating T cells from human donors with type 1 diabetes". *Nature Medicine* 22.12, pp. 1482–1487.

Baekkeskov, Steinunn et al. (1982). "Autoantibodies in newly diagnosed diabetic children immunoprecipitate human pancreatic islet cell proteins". *Nature* 298.5870, pp. 167–169.

Bakker, Paul IW de and Soumya Raychaudhuri (2012). "Interrogating the major histocompatibility complex with high-throughput genomics". *Human Molecular Genetics* 21.R1, R29–R36.

Balduzzi, Sara, Gerta Rücker, and Guido Schwarzer (2019). "How to perform a meta-analysis with R: a practical tutorial". *Evidence-Based Mental Health* 22, pp. 153–160.

Banting, Frederick Grant et al. (1922). "Pancreatic extracts in the treatment of diabetes mellitus". *Canadian Medical Association Journal* 12.3, p. 141.

Barker, Jennifer M et al. (2008). "Two single nucleotide polymorphisms identify the highest-risk diabetes HLA genotype: potential for rapid screening." *Diabetes* 57 (11), pp. 3152–5. ISSN: 1939-327X. DOI: 10.2337/db08-0605. URL: http://www.ncbi.nlm.nih.gov/pubmed/18694972http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2570414.

Barrett, Jeffrey C et al. (2009). "Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes". *Nature Genetics* 41.6, pp. 703–707. ISSN: 1061-4036. DOI: 10.1038/ng.381.

Bell, Graeme I, Shiro Horita, and John H Karam (1984). "A polymorphic locus near the human insulin gene is associated with insulin-dependent diabetes melliitus". *Diabetes* 33.2, pp. 176–183.

Bell, Graeme I, John H Karam, and William J Rutter (1981). "Polymorphic DNA region adjacent to the 5'end of the human insulin gene". *PNAS* 78.9, pp. 5759–5763.

Bell, Graeme I, Mark J Selby, and William J Rutter (1982). "The highly polymorphic region near the human insulin gene is composed of simple tandemly repeating sequences". *Nature* 295.5844, pp. 31–35.

Benjamini, Yoav and Daniel Yekutieli (2001). "The control of the false discovery rate in multiple testing under depencency". *The Annals of Statistics* 29.4, pp. 1165–1188. ISSN: 0090-5364. DOI: 10.1214/aos/1013699998.

Benner, Christian et al. (2016). "FINEMAP: efficient variable selection using summary data from genome-wide association studies". *Bioinformatics* 32.10, pp. 1493–1501.

Benner, Christian et al. (2017). "Prospects of Fine-Mapping Trait-Associated Genomic Regions by Using Summary Statistics from Genome-wide Association Stud-

ies." *American Journal of Human Genetics* 101.4, pp. 539–551. ISSN: 1537-6605. DOI: 10.1016/j.ajhg.2017.08.012. URL: http://www.ncbi.nlm.nih.gov/pubmed/28942963http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5630179.

Bennett, ST et al. (1995). "Susceptibility to human type 1 diabetes at IDDM2 is determined by tandem repeat variation at the insulin gene minisatellite locus". *Nature Genetics* 9.3, pp. 284–292.

Bingley, Polly J., David C. Boulware, and Jeffrey P. Krischer (2016). "The implications of autoantibodies to a single islet antigen in relatives with normal glucose tolerance: development of other autoantibodies and progression to type 1 diabetes". *Diabetologia* 59 (3), pp. 542–549. ISSN: 0012-186X. DOI: 10.1007/s00125-015-3830-2. URL: http://link.springer.com/10.1007/s00125-015-3830-2.

Bliss, Michael (1993). "The history of insulin". *Diabetes Care* 16.Supplement 3, pp. 4–7.

Bonifacio, Ezio et al. (2018). "Genetic scores to stratify risk of developing multiple islet autoantibodies and type 1 diabetes: a prospective study in children". *PLoS Medicine* 15.4, e1002548.

Botstein, David et al. (1980). "Construction of a genetic linkage map in man using restriction fragment length polymorphisms." *American Journal of Human Genetics* 32.3, p. 314.

Bottini, Nunzio et al. (2004). "A functional variant of lymphoid tyrosine phosphatase is associated with type I diabetes". *Nature Genetics* 36 (4), pp. 337–338. ISSN: 1061-4036. DOI: 10.1038/ng1323. URL: http://www.nature.com/doifinder/10.1038/ng1323.

Bottolo, Leonard and Sylvia Richardson (2010). "Evolutionary stochastic search for bayesian model exploration". *Bayesian Analysis* 5.3, pp. 583–618. ISSN: 19360975. DOI: 10.1214/10-BA523.

Bowman, Mark A, Edward H Leiter, and Mark A Atkinson (1994). "Prevention of diabetes in the NOD mouse: implications for therapeutic intervention in human disease". *Immunology Today* 15.3, pp. 115–120.

Boyle, Alan P et al. (2008). "High-resolution mapping and characterization of open chromatin across the genome". *Cell* 132.2, pp. 311–322.

Boyle, Alan P et al. (2012). "Annotation of functional variation in personal genomes using RegulomeDB". *Genome Research* 22, pp. 1790–1797. DOI: 10.1101/gr.137323.112..

Bradfield, Jonathan P et al. (2011). "A genome-wide meta-analysis of six type 1 diabetes cohorts identifies multiple associated loci." *PLoS Genetics* 7.9, e1002293. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1002293. URL: http://www.ncbi.nlm.nih.gov/pubmed/21980299http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3183083.

Breton, Marc D et al. (2020). "A randomized trial of closed-loop control in children with type 1 diabetes". *New England Journal of Medicine* 383.9, pp. 836–845.

Brorsson, Caroline A. et al. (2015). "Novel association between immune-mediated susceptibility loci and persistent autoantibody positivity in type 1 diabetes". *Diabetes* 64 (8), pp. 3017–3027. ISSN: 1939327X. DOI: 10.2337/db14-1730.

Brown, WM et al. (2009). "Overview of the MHC fine mapping data". *Diabetes, Obesity and Metabolism*.

Browning, Brian L and Sharon R Browning (2009). "A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals". *American Journal of Human Genetics* 84.2, pp. 210–223.

Browning, Brian L, Ying Zhou, and Sharon R Browning (2018). "A one-penny imputed genome from next-generation reference panels". *American Journal of Human Genetics* 103.3, pp. 338–348.

Buenrostro, Jason D et al. (2013). "Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position". *Nature Methods* 10.12, pp. 1213–1218.

Buenrostro, Jason D. et al. (2015). "ATAC-seq: A method for assaying chromatin accessibility genome-wide". *Current Protocols in Molecular Biology* January. ISSN: 19343647. DOI: 10.1002/0471142727.mb2129s109.

Buitinga, Mijke et al. (2018). "Inflammation-induced citrullinated glucose-regulated protein 78 elicits immune responses in human type 1 diabetes". *Diabetes* 67.11, pp. 2337–2348.

Bulik-Sullivan, Brendan K et al. (2015). "LD Score regression distinguishes confounding from polygenicity in genome-wide association studies". *Nature Genetics* 47.3, pp. 291–295.

Burren, Oliver S et al. (2017). "Chromosome contacts in activated T cells identify autoimmune disease candidate genes". *Genome Biology* 18.165, p. 165. DOI: 10.1186/s13059-017-1285-0.

Bycroft, Clare et al. (2018). "The UK Biobank resource with deep phenotyping and genomic data". *Nature* 562.7726, pp. 203–209.

Calderon, Diego et al. (2019). "Landscape of stimulation-responsive chromatin across diverse human immune cells". *Nature Genetics* 51.October, pp. 1494–1505. ISSN: 1546-1718. DOI: 10.1038/s41588-019-0505-9.

Cano-Gamez, Eddie and Gosia Trynka (2020). "From GWAS to function: using functional genomics to identify the mechanisms underlying complex diseases". *Frontiers in Genetics* 11, p. 424.

Chang, Christopher C et al. (2015). "Second-generation PLINK: rising to the challenge of larger and richer datasets". *Gigascience* 4.1, s13742–015.

Chen, Lu et al. (2016). "Genetic drivers of epigenetic and transcriptional variation in human immune cells". *Cell* 167.5, pp. 1398–1414.

Chiou, Joshua et al. (2021). "Interpreting type 1 diabetes risk with genetics and single-cell epigenomics". *Nature* 594.7863, pp. 398–402.

Chun, Sung et al. (2017). "Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types." *Nature Genetics* 49.4, pp. 600–605. ISSN: 1546-1718. DOI: 10.1038/ng.3795. URL: http://www.ncbi.nlm.nih.gov/pubmed/28218759http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5374036.

Chung, Wendy K et al. (2020). "Precision medicine in diabetes: a consensus report from the American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD)". *Diabetologia* 63.9, pp. 1671–1693.

Collins, FS et al. (2004). "Finishing the euchromatic sequence of the human genome". *Nature* 431.7011, pp. 931–945.

Cooper, Jason D et al. (2008). "Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci." *Nature Genetics* 40.12, pp. 1399–401. ISSN: 1546-1718. DOI: 10.1038/ng.249. URL: http://www.ncbi.nlm.nih.gov/pubmed/18978792http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2635556.

Cooper, Nicholas J et al. (2017). "Type 1 diabetes genome-wide association analysis with imputation identifies five new risk regions". *bioRxiv*, p. 120022.

Corces, M Ryan et al. (2017). "An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues". *Nature Methods* 14.10, pp. 959–962. DOI: 10.1038/nmeth.4396.

Cortes, Adrian et al. (2011). "Promise and pitfalls of the Immunochip". *Arthritis Research and Therapy* 13.1, p. 101. ISSN: 1478-6354. DOI: 10.1186/ar3204.

Cotsapas, Chris et al. (2011). "Pervasive Sharing of Genetic Effects in Autoimmune Disease". *PLoS Genetics* 7.8, e1002254. DOI: 10.1371/journal.pgen.1002254.

Crick, Francis (1970). "Central dogma of molecular biology". *Nature* 227.5258, pp. 561–563.

Crouch, Daniel JM et al. (2021). "Enhanced genetic analysis of type 1 diabetes by selecting variants on both effect size and significance, and by integration with autoimmune thyroid disease". *bioRxiv*.

Cucca, Francesco et al. (2001). "A correlation between the relative predisposition of MHC class II alleles to type 1 diabetes and the structure of their proteins." *Human Molecular Genetics* 10 (19), pp. 2025–37. ISSN: 0964-6906. DOI: 10.1093/hmg/10.19.2025. URL: http://www.ncbi.nlm.nih.gov/pubmed/11590120.

Dahl, Andy and Noah Zaitlen (2020). "Genetic Influences on Disease Subtypes". *Annual Review of Genomics and Human Genetics* 21, pp. 413–435.

Danila, Maria I et al. (2017). "Dense genotyping of immune-related regions identifies loci for rheumatoid arthritis risk and damage in African Americans". *Molecular Medicine* 23.1, pp. 177–187.

Danko, Charles G et al. (2018). "Dynamic evolution of regulatory element ensembles in primate CD4+ T cells." *Nature Ecology & Evolution* 2.3, pp. 537–548. ISSN: 2397-334X. DOI: 10.1038/s41559-017-0447-5. URL: http://www.ncbi.nlm.nih.gov/pubmed/29379187http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5957490.

Das, Sayantan et al. (2016). "Next-generation genotype imputation service and methods". *Nature Genetics* 48.10, pp. 1284–1287. DOI: 10.1038/ng.3656.

Davis, James P et al. (2019). "Enhancer deletion and allelic effects define a regulatory molecular mechanism at the VLDLR cholesterol GWAS locus". *Human Molecular Genetics* 28.6, pp. 888–895.

de Jong, V et al. (2016). "Variation in the CTLA4 3'UTR has phenotypic consequences for autoreactive T cells and associates with genetic risk for type 1 diabetes". *Genes and Immunity* 17 (1), pp. 75–78. ISSN: 1466-4879. DOI: 10.1038/

gene.2015.51. URL: http://www.ncbi.nlm.nih.gov/pubmed/26656450http://www.nature.com/doifinder/10.1038/gene.2015.51.

DeLong, Elizabeth R, David M DeLong, and Daniel L Clarke-Pearson (1988). "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach". *Biometrics*, pp. 837–845.

Delong, T. et al. (2016). "Pathogenic CD4 T cells in type 1 diabetes recognize epitopes formed by peptide fusion". *Science* 351 (6274), pp. 711–714. ISSN: 0036-8075. DOI: 10.1126/science.aad2791. URL: http://www.sciencemag.org/cgi/doi/10.1126/science.aad2791.

Dooley, James et al. (2016). "Genetic predisposition for beta cell fragility underlies type 1 and type 2 diabetes". *Nature Genetics* 48.5, pp. 519–527. DOI: 10.1038/ng.3531.

Erlich, H (2012). "HLA DNA typing: past, present, and future". *Tissue Antigens* 80.1, pp. 1–11.

Erlich, Henry et al. (2008). "HLA DR-DQ haplotypes and genotypes and type 1 diabetes risk: analysis of the type 1 diabetes genetics consortium families". *Diabetes* 57.4, pp. 1084–1092.

Ernst, Jason and Manolis Kellis (2017). "Chromatin-state discovery and genome annotation with ChromHMM". *Nature Protocols* 12.12, pp. 2478–2492. ISSN: 1754-2189. DOI: 10.1038/nprot.2017.124.

Excoffier, Laurent and Montgomery Slatkin (1995). "Maximum-Likelihood Estimation of Molecular Haplotype Frequencies in a Diploid Population". *Molecular Biology and Evolution* 12.5, pp. 921–927.

Faegan, Brian G et al. (2018). "Risankizumab in patients with moderate to severe Crohn's disease: an open-label extension study". *Lancet Gastroenterology & Hepatology* 3.10, pp. 671–680.

Fairfax, Benjamin P et al. (2012). "Genetics of gene expression in primary immune cells identifies cell type – specific master regulators and roles of HLA alleles". *Nature Genetics* 44.5, pp. 502–510. ISSN: 1061-4036. DOI: 10.1038/ng.2205.

Fairfax, Benjamin P et al. (2014). "Innate Immune Activity Conditions the Effect of Regulatory Variants upon Monocyte Gene Expression". *Science* 343.3, p. 1246949. DOI: 10.1126/science.1246949.

Fang, Hai et al. (2019). "A genetics-led approach defines the drug target landscape of 30 immune-related traits". *Nature Genetics* 51.7, pp. 1082–1091. DOI: 10.1038/s41588-019-0456-1.

Farh, Kyle Kai-How et al. (2015). "Genetic and epigenetic fine mapping of causal autoimmune disease variants". *Nature* 518.7539, pp. 337–343.

Fasolino, Maria et al. (2021). "Multiomics single-cell analysis of human pancreatic islets reveals novel cellular states in health and type 1 diabetes". *bioRxiv*.

Ferrat, Lauric A et al. (2020). "A combined risk score enhances prediction of type 1 diabetes among susceptible children". *Nature Medicine* 26.8, pp. 1247–1255.

Ferreira, Ricardo C. et al. (2013). "Functional IL6R 358Ala Allele Impairs Classical IL-6 Receptor Signaling and Influences Risk of Diverse Inflammatory Diseases". *PLoS Genetics* 9.4, e1003444. ISSN: 15537390. DOI: 10.1371/journal.pgen.1003444.

Fort, Alexandre et al. (2017). "MBV: a method to solve sample mislabeling and detect technical bias in large combined genotype and sequencing assay datasets". *Bioinformatics* 33.12, pp. 1895–1897. DOI: 10.1093/bioinformatics/btx074.

Fotiadou, Christina et al. (2018). "Targeting IL-23 in psoriasis: current perspectives". *Psoriasis: Targets and Therapy* 8, pp. 1–5. DOI: 10.2147/PTT.S98893.

Fulco, Charles P et al. (2016). "Systematic mapping of functional enhancer–promoter connections with CRISPR interference". *Science* 354.6313, pp. 769–773.

Gaglia, Jason and Stephan Kissler (2019). "Anti-CD3 Antibody for the Prevention of Type 1 Diabetes: A Story of Perseverance". *Biochemistry* 58, pp. 4107–4111. DOI: 10.1021/acs.biochem.9b00707.

Gaspar, John M (2018). "Improved peak-calling with MACS2". *bioRxiv*, pp. 1–16.

Gaziano, John Michael et al. (2016). "Million Veteran Program: A mega-biobank to study genetic influences on health and disease". *Journal of Clinical Epidemiology* 70, pp. 214–223.

Ge, Yan et al. (2016). "Targeted Deep Sequencing in Multiple-Affected Sibships of European Ancestry Identifies Rare Deleterious Variants in PTPN22 That Confer Risk for Type 1 Diabetes." *Diabetes* 65 (3), pp. 794–802. ISSN: 1939-327X. DOI: 10.2337/db15-0322. URL: http://www.ncbi.nlm.nih.gov/pubmed/26631741http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4764149.

Ge, Yan et al. (2017). "UBASH3A mediates risk for type 1 diabetes through inhibition of T-cell receptor–induced NF-κB signaling". *Diabetes* 66.7, pp. 2033–2043.

Ge, Yan et al. (2019). "UBASH3A regulates the synthesis and dynamics of TCR–CD3 complexes". *The Journal of Immunology* 203.11, pp. 2827–2836.

Giambartolomei, Claudia et al. (2014). "Bayesian test for colocalisation between pairs of genetic association studies using summary statistics." *PLoS Genetics* 10.5, e1004383. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1004383.

Goeddel, David V et al. (1979). "Expression in Escherichia coli of chemically synthesized genes for human insulin". *PNAS* 76.1, pp. 106–110.

Gorman, Jacquelyn A et al. (2017). "The A946T variant of the RNA sensor IFIH1 mediates an interferon program that limits viral infection but increases the risk for autoimmunity". *Nature Immunology* 18 (7), pp. 744–752. ISSN: 1529-2908. DOI: 10.1038/ni.3766. URL: http://www.nature.com/doifinder/10.1038/ni.3766.

Grant, Struan F A et al. (2009). "Follow-up analysis of genome-wide association data identifies novel loci for type 1 diabetes." *Diabetes* 58.1, pp. 290–5. ISSN: 1939-327X. DOI: 10.2337/db08-1022. URL: http://www.ncbi.nlm.nih.gov/pubmed/18840781http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2606889.

Gruessner, Rainer WG and Angelika C Gruessner (2013). "The current state of pancreas transplantation". *Nature Reviews Endocrinology* 9.9, pp. 555–562.

GTEx Consortium (2020). "The GTEx Consortium atlas of genetic regulatory effects across human tissues". *Science* 369.6509, pp. 1318–1330.

Gutierrez-Arcelus, Maria et al. (2020). "Allele-specific expression changes dynamically during T cell activation in HLA and other autoimmune loci". *Nature Genetics* 52.3, pp. 247–253.

Hakonarson, Hakon et al. (2008). "A novel susceptibility locus for type 1 diabetes on Chr12q13 identified by a genome-wide association study." *Diabetes* 57.4, pp. 1143–6. ISSN: 1939-327X. DOI: 10.2337/db07-1305. URL: http://www.ncbi.nlm.nih.gov/pubmed/18198356.

Harrow, Jennifer et al. (2012). "GENCODE: The reference human genome annotation for The ENCODE Project". *Genome Research* 22, pp. 1760–1774. DOI: 10.1101/gr.135350.111..

Herold, Kevan C et al. (2019). "An anti-CD3 antibody, teplizumab, in relatives at risk for type 1 diabetes". *New England Journal of Medicine* 381.7, pp. 603–613.

Hippich, Markus et al. (2019). "Genetic contribution to the divergence in type 1 diabetes risk between children from the general population and children from affected families". *Diabetes* 68.4, pp. 847–857.

Hormozdiari, Farhad et al. (2014). "Identifying causal variants at loci with multiple signals of association". *Genetics* 198.2, pp. 497–508.

Howie, Bryan et al. (2012). "Fast and accurate genotype imputation in genome-wide association studies through pre-phasing". *Nature Genetics* 44.8, pp. 955–959.

Howson, JMM et al. (2013). "HLA class II gene associations in African American Type 1 diabetes reveal a protective HLA-DRB1* 03 haplotype". *Diabetic Medicine* 30.6, pp. 710–716.

Hu, Xinli et al. (2015). "Additive and interaction effects at three amino acid positions in HLA-DQ and HLA-DR molecules drive type 1 diabetes risk". *Nature Genetics* 47.8, pp. 898–905.

Huang, Jie et al. (2012). "1000 Genomes-based imputation identifies novel and refined associations for the Wellcome Trust Case Control Consortium phase 1 Data." *European Journal of Human Genetics* 20.7, pp. 801–5. ISSN: 1476-5438. DOI: 10.1038/ejhg.2012.3. URL: http://www.ncbi.nlm.nih.gov/pubmed/22293688http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3376268.

Hukku, Abhay et al. (2020). "Probabilistic Colocalization of Genetic Variants from Complex and Molecular Traits: Promise and Limitations". *bioRxiv*, p. 2020.07.01.182097. DOI: 10.1101/2020.07.01.182097. URL: https://doi.org/10.1101/2020.07.01.182097.

Ilonen, J. et al. (2016). "Genetic susceptibility to type 1 diabetes in childhood – estimation of HLA class II associated disease risk and class II effect in various phases of islet autoimmunity". *Pediatric Diabetes* 17, pp. 8–16. ISSN: 13995448. DOI: 10.1111/pedi.12327.

Ilonen, Jorma et al. (2017). "Primary islet autoantibody at initial seroconversion and autoantibodies at diagnosis of type 1 diabetes as markers of disease heterogeneity". *Pediatric Diabetes*. ISSN: 1399543X. DOI: 10.1111/pedi.12545. URL: http://doi.wiley.com/10.1111/pedi.12545.

Insel, Richard A et al. (2015). "Staging presymptomatic type 1 diabetes: a scientific statement of JDRF, the Endocrine Society, and the American Diabetes Association". *Diabetes Care* 38.10, pp. 1964–1974.

Inshaw, Jamie R. J. et al. (2017). "The chromosome 6q22.33 region is associated with age at diagnosis of type 1 diabetes and disease risk in those diagnosed under 5 years of age". *Diabetologia*, pp. 1–11. ISSN: 0012-186X. DOI: 10.1007/s00125-017-4440-y. URL: http://link.springer.com/10.1007/s00125-017-4440-y.

Inshaw, Jamie RJ et al. (2020). "Genetic variants predisposing most strongly to type 1 diabetes diagnosed under age 7 years lie near candidate genes that function in the immune system and in pancreatic β-cells". *Diabetes Care* 43.1, pp. 169–177.

Ise, Wataru et al. (2010). "CTLA-4 suppresses the pathogenicity of self antigen–specific T cells by cell-intrinsic and cell-extrinsic mechanisms". *Nature Immunology* 11 (2), pp. 129–135. ISSN: 1529-2908. DOI: 10.1038/ni.1835. URL: http://www.nature.com/doifinder/10.1038/ni.1835.

Isobe, Noriko et al. (2015). "An ImmunoChip study of multiple sclerosis risk in African Americans". *Brain* 138.6, pp. 1518–1530.

Jakkula, Eveliina et al. (2008). "The genome-wide patterns of variation expose significant substructure in a founder population". *American Journal of Human Genetics* 83.6, pp. 787–794.

Javierre, Biola M. et al. (2016). "Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters". *Cell* 167.5, pp. 1369–1384. ISSN: 00928674. DOI: 10.1016/j.cell.2016.09.037.

Jia, Xiaoming et al. (2013). "Imputing amino acid polymorphisms in human leukocyte antigens". *PloS One* 8.6, e64683.

Jiang, Hongshan et al. (2014). "Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads". *BMC Bioinformatics* 15.182.

Jonsson, Malin K.B. et al. (2016). "A Transcriptomic and Epigenomic Comparison of Fetal and Adult Human Cardiac Fibroblasts Reveals Novel Key Transcription Factors in Adult Cardiac Fibroblasts". *JACC: Basic to Translational Science* 1.7, pp. 590–602.

Jostins, Luke (2013). "Using next-generation genomic datasets in disease association". PhD thesis. University of Cambridge.

Kaprio, J et al. (1992). "Concordance for type 1 (insulin-dependent) and type 2 (non-insulin-dependent) diabetes mellitus in a population-based cohort of twins in Finland". *Diabetologia* 35.11, pp. 1060–1067.

Karnes, Jason H et al. (2017). "Comparison of HLA allelic imputation programs". *PloS One* 12.2, e0172444.

Kasela, Silva et al. (2017). "Pathogenic implications for autoimmune mechanisms derived by comparative eQTL analysis of CD4+versus CD8+T cells". *PLoS Genetics* 13.3, e1006643. ISSN: 15537404. DOI: 10.1371/journal.pgen.1006643.

Kazeem, G R and M Farrall (2005). "Integrating Case-control and TDT Studies". *Annals of Human Genetics* 69, pp. 329–335. DOI: 10.1046/j.1529-8817.2005.00156.x.

Keen, Harry et al. (1980). "Human insulin produced by recombinant DNA technology: safety and hypoglycaemic potency in healthy men". *The Lancet* 316.8191, pp. 398–401.

Khera, Amit V and Sekar Kathiresan (2017). "Is coronary atherosclerosis one disease or many? Setting realistic expectations for precision medicine". *Circulation* 135.11, pp. 1005–1007.

Kichaev, Gleb and Bogdan Pasaniuc (2015). "Leveraging Functional-Annotation Data in Trans-ethnic Fine-Mapping Studies". *American Journal of Human Genetics* 97.2, pp. 260–271. ISSN: 15376605. DOI: 10.1016/j.ajhg.2015.06.007.

King, Emily A, J Wade Davis, and Jacob F Degner (2019). "Are drug targets with genetic support twice as likely to be approved ? Revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval". *PLoS Genetics* 15.12. DOI: 10.1371/journal.pgen.1008489.

Klemm, Sandy L, Zohar Shipony, and William J Greenleaf (2019). "Chromatin accessibility and the regulatory epigenome". *Nature Reviews Genetics* 20.4, pp. 207–220.

Kracht, Maria J L et al. (2017). "Autoimmunity against a defective ribosomal insulin gene product in type 1 diabetes". *Nature Medicine* 23 (4), pp. 501–507. ISSN: 1078-8956. DOI: 10.1038/nm.4289. URL: http://www.nature.com/doifinder/10.1038/nm.4289.

Krischer, Jeffrey P. et al. (2015). "The 6 year incidence of diabetes-associated autoantibodies in genetically at-risk children: the TEDDY study". *Diabetologia* 58 (5), pp. 980–987. ISSN: 14320428. DOI: 10.1007/s00125-015-3514-y.

Krischer, Jeffrey P et al. (2017). "The influence of type 1 diabetes genetic susceptibility regions, age, sex, and family history on the progression from multiple autoantibodies to type 1 diabetes: a TEDDY study report". *Diabetes* 66.12, pp. 3122–3129.

Krischer, Jeffrey P et al. (2019). "Predicting Islet Cell Autoimmunity and Type 1 Diabetes : An 8-Year TEDDY Study Progress Report". *Diabetes Care* 42.June, pp. 1051–1060. DOI: 10.2337/dc18-2282.

Kundaje, Anshul et al. (2015). "Integrative analysis of 111 reference human epigenomes". *Nature* 518.7539, pp. 317–330.

Kundu, Kousik et al. (2020). "Genetic associations at regulatory phenotypes improve fine-mapping of causal variants for twelve immune-mediated diseases". *bioRxiv*, p. 2020.01.15.907436. DOI: 10.1101/2020.01.15.907436. URL: https://doi.org/10.1101/2020.01.15.907436.

Ladner, Martha B et al. (2005). "Association of the single nucleotide polymorphism C1858T of the PTPN22 gene with type 1 diabetes". *Human immunology* 66.1, pp. 60–64.

LaFramboise, Thomas (2009). "Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances". *Nucleic Acids Research* 37.13, pp. 4181–4193.

Lam, Max et al. (2019). "Comparative genetic architectures of schizophrenia in East Asian and European populations". *Nature Genetics* 51.12, pp. 1670–1678.

Langmead, Ben and Steven L Salzberg (2012). "Fast gapped-read alignment with Bowtie 2". *Nature Methods* 9.4, pp. 357–359. DOI: 10.1038/nmeth.1923.

Lawrence, Michael, Robert Gentleman, and Vincent Carey (2009). "rtracklayer: an R package for interfacing with genome browsers". *Bioinformatics* 25, pp. 1841–1842. DOI: 10.1093/bioinformatics/btp328. URL: http://bioinformatics.oxfordjournals.org/content/25/14/1841.abstract.

Lawrence, Michael et al. (2013). "Software for Computing and Annotating Genomic Ranges". *PLoS Computational Biology* 9 (8). DOI: 10.1371/journal.pcbi.1003118. URL: http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1003118.

Lee, David, Oliver Redfern, and Christine Orengo (2007). "Predicting protein function from sequence and structure". *Nature Reviews Molecular Cell Biology* 8.12, pp. 995–1005.

Leete, Pia et al. (2016). "Differential insulitic profiles determine the extent of β-cell destruction and the age at onset of type 1 diabetes". *Diabetes* 65.5, pp. 1362–1369.

Li, Heng (2018). "Minimap2: Pairwise alignment for nucleotide sequences". *Bioinformatics* 34.18, pp. 3094–3100. ISSN: 14602059. DOI: 10.1093/bioinformatics/bty191.

Li, Na and Matthew Stephens (2003). "Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data". *Genetics* 165.4, pp. 2213–2233.

Li, Qunhua et al. (2011). "Measuring reproducibility of high-throughput experiments". *Annals of Applied Statistics* 5.3, pp. 1752–1779. ISSN: 19326157. DOI: 10.1214/11-AOAS466.

Li, Xinrui et al. (2013). "Allelic-dependent expression of an activating Fc receptor on B cells enhances humoral immune responses". *Science Translational Medicine* 5.216, 216ra175–216ra175.

Liao, Yang, Gordon K. Smyth, and Wei Shi (2014). "FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features". *Bioinformatics* 30.7, pp. 923–930. ISSN: 14602059. DOI: 10.1093/bioinformatics/btt656.

Lieberman-Aiden, Erez et al. (2009). "Comprehensive mapping of long-range interactions reveals folding principles of the human genome". *Science* 326.5950, pp. 289–293.

Little, Alicia J et al. (2015). "The mechanism of V (D) J recombination". *Molecular Biology of B cells*. Elsevier, pp. 13–34.

Liu, Boxiang et al. (2019). "Abundant associations with gene expression complicate GWAS follow-up". *Nature Genetics* 51.5, pp. 768–769. ISSN: 1061-4036. DOI: 10.1038/s41588-019-0404-0. URL: http://www.nature.com/articles/s41588-019-0404-0.

Lizio, Marina et al. (2015). "Gateways to the FANTOM5 promoter level mammalian expression atlas." *Genome Biology* 16, p. 22. ISSN: 1474-760X. DOI: 10.1186/s13059-014-0560-6. URL: http://www.ncbi.nlm.nih.gov/pubmed/25723102http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4310165.

Loh, Po-ru et al. (2016). "Reference-based phasing using the Haplotype Reference Consortium panel". *Nature Genetics* 48.11, pp. 1443–1448. DOI: 10.1038/ng.3679.

Lönnrot, Maria et al. (2017). "Respiratory infections are temporally associated with initiation of type 1 diabetes autoimmunity: the TEDDY study". *Diabetologia* 60.10, pp. 1931–1940.

Loos, Ruth JF (2018). "The genetics of adiposity". *Current Opinion in Genetics & Development* 50, pp. 86–95.

— (2020). "15 years of genome-wide association studies and no signs of slowing down". *Nature Communications* 11.1, pp. 1–3.

Love, Michael I., Wolfgang Huber, and Simon Anders (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". *Genome Biology* 15 (12), p. 550. DOI: 10.1186/s13059-014-0550-8.

Luo, Yang et al. (2020). "A high-resolution HLA reference panel capturing global population diversity enables multi-ethnic fine-mapping in HIV host response". *medRxiv*.

Maller, Julian B et al. (2012). "Bayesian refinement of association signals for 14 loci in 3 common diseases". *Nature Genetics* 44.12, pp. 1294–1301.

Manichaikul, Ani et al. (2010). "Robust relationship inference in genome-wide association studies". *Bioinformatics* 26.22, pp. 2867–2873.

Manolio, Teri A et al. (2009). "Finding the missing heritability of complex diseases". *Nature* 461.7265, pp. 747–753.

Marchini, Jonathan and Bryan Howie (2010). "Genotype imputation for genome-wide association studies". *Nature Reviews Genetics* 11.7, pp. 499–511. ISSN: 1471-0064. DOI: 10.1038/nrg2796.

Marroqui, Laura et al. (2015). "TYK2, a candidate gene for type 1 diabetes, modulates apoptosis and the innate immune response in human pancreatic β-cells". *Diabetes* 64.11, pp. 3808–3817.

Martin, Alicia R et al. (2019). "Clinical use of current polygenic risk scores may exacerbate health disparities". *Nature Genetics* 51.4, pp. 584–591.

Maurano, Matthew T et al. (2012). "Systematic localization of common disease-associated variation in regulatory DNA". *Science* 337.6099, pp. 1190–1195.

Mayer-Davis, Elizabeth J et al. (2017). "Incidence trends of type 1 and type 2 diabetes among youths, 2002–2012". *New England Journal of Medicine* 376, pp. 1419–1429.

Mayor, Neema P et al. (2015). "HLA typing for the next generation". *PloS One* 10.5, e0127153.

McEvoy, Brian P et al. (2011). "Human population dispersal "Out of Africa" estimated from linkage disequilibrium and allele frequencies of SNPs". *Genome Research* 21.6, pp. 821–829.

McGinnis, Ralph, Sagiv Shifman, and Ariel Darvasi (2002). "Power and efficiency of the TDT and case-control design for association scans". *Behavior Genetics* 32.2, pp. 135–144.

McGinty, John W et al. (2014). "Recognition of posttranslationally modified GAD65 epitopes in subjects with type 1 diabetes". *Diabetes* 63.9, pp. 3033–3040.

McPherson, John D et al. (2001). "A physical map of the human genome". *Nature* 409.6822, pp. 934–941.

Melnikov, Alexandre et al. (2012). "Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay". *Nature Biotechnology* 30.3, pp. 271–277.

Mendel, Gregor (1865). "Experiments in plant hybridization (1865)". *Verhandlungen des naturforschenden Vereins Brünn.) Available online: www. mendelweb. org/Mendel. html (accessed on 1 January 2013).*

Mitchell, Maria K et al. (2004). "The New York Cancer Project: rationale, organization, design, and baseline characteristics". *Journal of Urban Health* 81.2, pp. 301–310.

Moore, Jill E et al. (2020). "Expanded encyclopaedias of DNA elements in the human and mouse genomes". *Nature* 583.7818, pp. 699–710.

Morgan, Noel G and Sarah J Richardson (2014). "Enteroviruses as causative agents in type 1 diabetes: loose ends or lost cause?" *Trends in Endocrinology & Metabolism* 25.12, pp. 611–619.

Morgan, Noel G et al. (2014). "Islet inflammation in human type 1 diabetes mellitus". *IUBMB life* 66.11, pp. 723–734.

Morris, John A et al. (2021). "Discovery of target genes and pathways of blood trait loci using pooled CRISPR screens and single cell RNA sequencing". *bioRxiv*.

Mueller, Patricia W et al. (2006). "Genetics of Kidneys in Diabetes (GoKinD) study: a genetics collection available for identifying genetic susceptibility factors for diabetic nephropathy in type 1 diabetes". *Journal of the American Society of Nephrology* 17.7, pp. 1782–1790.

Murphy, Kenneth, Paul Travers, and M Walport (2012). *Janeway's immunobiology 8th edition, Garland Science*.

Musunuru, Kiran et al. (2010). "From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus". *Nature* 466.7307, pp. 714–719.

Mychaleckyj, Josyf C et al. (2010). "HLA genotyping in the international Type 1 Diabetes Genetics Consortium". *Clinical Trials* 7.1_suppl, S75–S87.

Naito, Tatsuhiko et al. (2021). "A deep learning method for HLA imputation and trans-ethnic MHC fine-mapping of type 1 diabetes". *Nature Communications* 12.1, pp. 1–14.

Nasrallah, Rabab et al. (2020). "A distal enhancer at risk locus 11q13. 5 promotes suppression of colitis by T reg cells". *Nature* 583.7816, pp. 447–452.

Neerincx, Andreas et al. (2013). "NLRC5, at the heart of antigen presentation". *Frontiers in Immunology* 4, p. 397.

Nejentsev, S. et al. (2009). "Rare Variants of IFIH1, a Gene Implicated in Antiviral Responses, Protect Against Type 1 Diabetes". *Science* 324 (5925), pp. 387–389. ISSN: 0036-8075. DOI: 10.1126/science.1167728. URL: http://www.sciencemag.org/cgi/doi/10.1126/science.1167728.

Nejentsev, Sergey et al. (2007). "Localization of type 1 diabetes susceptibility to the MHC class I genes HLA-B and HLA-A". *Nature* 450 (7171), pp. 887–892. ISSN: 0028-0836. DOI: 10.1038/nature06406. URL: http://www.nature.com/doifinder/10.1038/nature06406.

Nelson, Matthew R et al. (2015). "The support of human genetic evidence for approved drug indications". *Nature Genetics* 47.8, pp. 856–860. ISSN: 1061-4036. DOI: 10.1038/ng.3314.

Neph, Shane et al. (2012). "BEDOPS : high-performance genomic feature operations". *Bioinformatics* 28.14, pp. 1919–1920. DOI: 10.1093/bioinformatics/bts277.

Nerup, J et al. (1974). "HL-A antigens and diabetes mellitus". *The Lancet* 304.7885, pp. 864–866.

Neuwirth, Erich (2014). *RColorBrewer: ColorBrewer Palettes*. R package version 1.1-2. URL: https://CRAN.R-project.org/package=RColorBrewer.

Nisticò, Lorenza et al. (1996). "The CTLA-4 gene region of chromosome 2q33 is linked to, and associated with, type 1 diabetes". *Human Molecular Genetics* 5.7, pp. 1075–1080.

Noble, Janelle A et al. (2010). "HLA class I and genetic susceptibility to type 1 diabetes: results from the Type 1 Diabetes Genetics Consortium". *Diabetes* 59.11, pp. 2972–2979.

Noble, Janelle A et al. (2013). "HLA class II genotyping of African American type 1 diabetic patients reveals associations unique to African haplotypes". *Diabetes* 62.9, pp. 3292–3299.

Okada, Yukinori et al. (2014). "Genetics of rheumatoid arthritis contributes to biology and drug discovery". *Nature* 506, pp. 376–381. ISSN: 0028-0836. DOI: 10.1038/nature12873.

Onengut-Gumuscu, Suna, Jane H. Buckner, and Patrick Concannon (2006). "A haplotype-based analysis of the PTPN22 locus in type 1 diabetes". *Diabetes* 55 (10), pp. 2883–2889. ISSN: 00121797. DOI: 10.2337/db06-0225.

Onengut-Gumuscu, Suna et al. (2015). "Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers". *Nature Genetics* 47.4, pp. 381–386.

Onengut-Gumuscu, Suna et al. (2019). "Type 1 diabetes risk in African-ancestry participants and utility of an ancestry-specific genetic risk score". *Diabetes Care* 42.3, pp. 406–415.

Onengut-Gumuscu, Suna et al. (2020). "Novel genetic risk factors influence progression of islet autoimmunity to type 1 diabetes". *Scientific Reports* 10.1, pp. 1–7.

Ooi, Joshua D. et al. (2017). "Dominant protection from HLA-linked autoimmunity by antigen-specific regulatory T cells". *Nature* 545 (7653), pp. 243–247. ISSN: 0028-0836. DOI: 10.1038/nature22329. URL: http://www.nature.com/doifinder/10.1038/nature22329.

Oram, Richard A et al. (2016). "A type 1 diabetes genetic risk score can aid discrimination between type 1 and type 2 diabetes in young adults". *Diabetes Care* 39.3, pp. 337–344.

Orban, Tihamer et al. (2011). "Co-stimulation modulation with abatacept in patients with recent-onset type 1 diabetes: a randomised, double-blind, placebo-controlled

trial". *The Lancet* 378 (9789), pp. 412–419. ISSN: 0140-6736. DOI: 10.1016/S0140-6736(11)60886-6. URL: http://www.sciencedirect.com/science/article/pii/S0140673611608866?via%3Dihub.

Pagliuca, Felicia W et al. (2014). "Generation of functional human pancreatic β cells in vitro". *Cell* 159.2, pp. 428–439.

Pan, Yidan et al. (2020). "Fine-mapping within eQTL credible intervals by expression CROP-seq". *Biology Methods and Protocols* 5.1, bpaa008.

Paola Benaglio, Jacklyn Newsome, Jee Yun Han, Joshua Chiou, Anthony Aylward, Sierra Corban, Mei-Lin Okino, Jaspreet Kaur, David U Gorkin, Kyle J Gaulton (2020). "Mapping genetic effects on cell type-specific chromatin accessibility and annotating complex trait variants using single nucleus ATAC-seq". *bioRxiv*. DOI: doi:https://doi.org/10.1101/2020.12.03.387894.

Park, Peter J (2009). "ChIP–seq: advantages and challenges of a maturing technology". *Nature Reviews Genetics* 10.10, pp. 669–680.

Patel, Kashyap A et al. (2016). "Type 1 diabetes genetic risk score: a novel tool to discriminate monogenic and type 1 diabetes". *Diabetes* 65.7, pp. 2094–2099.

Patterson, Christopher C et al. (2009). "Incidence trends for childhood type 1 diabetes in Europe during 1989–2003 and predicted new cases 2005–20: a multicentre prospective registration study". *The Lancet* 373.9680, pp. 2027–2033.

Patterson, Christopher C et al. (2019). "Worldwide estimates of incidence, prevalence and mortality of type 1 diabetes in children and adolescents: Results from the International Diabetes Federation Diabetes Atlas". *Diabetes Research and Clinical Practice* 157, p. 107842.

Price, Alkes L et al. (2008). "Long-range LD can confound genome scans in admixed populations". *American Journal of Human Genetics* 83.1, p. 132.

Privé, Florian et al. (2021). "High-resolution portability of 245 polygenic scores when derived and applied in the same cohort". *medRxiv*.

Pugliese, Alberto et al. (1997). "The insulin gene is transcribed in the human thymus and transcription levels correlate with allelic variation at the INS VNTR-IDDM2 susceptibility locus for type 1 diabetes". *Nature Genetics* 15 (3), pp. 293–297. ISSN: 1061-4036. DOI: 10.1038/ng0397-293. URL: http://www.nature.com/doifinder/10.1038/ng0397-293.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Raab, Jennifer et al. (2016). "Capillary blood islet autoantibody screening for identifying pre-type 1 diabetes in the general population: design and initial results of the Fr1da study". *BMJ Open* 6.5, e011144.

Rama, Ines and Josep M Grinyo (2010). "Malignancy after renal transplantation: the role of immunosuppression". *Nature Reviews Nephrology* 6.9, pp. 511–519.

Ramírez, Fidel et al. (2016). "deepTools2: a next generation web server for deep-sequencing data analysis". *Nucleic Acids Research* 44.W1, W160–W165.

Ramos-rodríguez, Mireia et al. (2019). "The impact of proinflammatory cytokines on the beta-cell regulatory landscape provides insights into the genetics of type 1 diabetes". *Nature Genetics* 51.November, pp. 1588–1595. ISSN: 1546-1718. DOI: 10.1038/s41588-019-0524-6.

Redondo, Maria J et al. (2008). "Concordance for islet autoimmunity among monozygotic twins". *New England Journal of Medicine* 359.26, pp. 2849–2850.

Redondo, Maria J et al. (2017). "TCF7L2 Genetic Variants Contribute to Phenotypic Heterogeneity of Type 1 Diabetes." *Diabetes Care*, p. dc170961. ISSN: 1935-5548. DOI: 10.2337/dc17-0961. URL: http://www.ncbi.nlm.nih.gov/pubmed/29025879.

Regier, Allison A et al. (2018). "Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects".

Nature Communications 9.4038. ISSN: 2041-1723. DOI: 10.1038/s41467-018-06159-4.

Rewers, M et al. (1996). "Newborn screening for HLA markers associated with IDDM: diabetes autoimmunity study in the young (DAISY)". *Diabetologia* 39.7, pp. 807–812.

Rewers, Marian and Johnny Ludvigsson (2016). "Environmental risk factors for type 1 diabetes". *The Lancet* 387, pp. 2340–2348.

Rich, SS, LR Weitkamp, and J Barbosa (1984). "Genetic heterogeneity of insulin-dependent (type I) diabetes mellitus: evidence from a study of extended haplotypes." *American Journal of Human Genetics* 36.5, p. 1015.

Rich, Stephen S (2017). "Genetics and its potential to improve T1D care". *Current Opinion in Endocrinology, Diabetes, and Obesity* 24.4, p. 279.

Rich, Stephen S et al. (2006). "The type 1 diabetes genetics consortium". *Annals of the New York Academy of Sciences* 1079.1, pp. 1–8.

Richardson, Sarah J et al. (2009). "The prevalence of enteroviral capsid protein vp1 immunostaining in pancreatic islets in human type 1 diabetes". *Diabetologia* 52.6, pp. 1143–1151.

Rieck, M. et al. (2007). "Genetic Variation in PTPN22 Corresponds to Altered Function of T and B Lymphocytes". *The Journal of Immunology* 179 (7), pp. 4704–4710. ISSN: 0022-1767. DOI: 10.4049/jimmunol.179.7.4704. URL: http://www.jimmunol.org/cgi/doi/10.4049/jimmunol.179.7.4704.

Robinson, James et al. (2020). "Ipd-imgt/hla database". *Nucleic Acids Research* 48.D1, pp. D948–D955.

Robinson, Mark D, Davis J Mccarthy, and Gordon K Smyth (2010). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data". *Bioinformatics* 26.1, pp. 139–140. DOI: 10.1093/bioinformatics/btp616.

Robinson, Mark D and Alicia Oshlack (2010). "A scaling normalization method for differential expression analysis of RNA-seq data". *Genome Biology* 11.3, R25.

Roep, Bart O et al. (2021). "Type 1 diabetes mellitus as a disease of the β-cell (do not blame the immune system?)" *Nature Reviews Endocrinology* 17.3, pp. 150–161.

Roychoudhuri, Rahul et al. (2016). "BACH2 regulates CD8+ T cell differentiation by controlling access of AP-1 factors to enhancers". *Nature Immunology* 17.7, pp. 851–860. DOI: 10.1038/ni.3441.

Rubinacci, Simone, Olivier Delaneau, and Jonathan Marchini (2020). "Genotype imputation using the positional burrows wheeler transform". *PLoS Genetics* 16.11, e1009049.

Russell, Steven J et al. (2014). "Outpatient glycemic control with a bionic pancreas in type 1 diabetes". *New England Journal of Medicine* 371.4, pp. 313–325.

San Luis, Boris et al. (2011). "Sts-2 is a phosphatase that negatively regulates zeta-associated protein (ZAP)-70 and T cell receptor signaling pathways". *Journal of Biological Chemistry* 286.18, pp. 15943–15954.

Sandborn, William J et al. (2017). "Tofacitinib as Induction and Maintenance Therapy for Ulcerative Colitis". *New England Journal of Medicine* 376.18, pp. 1723–1736. DOI: 10.1056/NEJMoa1606910.

Scally, Stephen W et al. (2013). "A molecular basis for the association of the HLA-DRB1 locus, citrullination, and rheumatoid arthritis." *The Journal of Experimental Medicine* 210 (12), pp. 2569–82. ISSN: 1540-9538. DOI: 10.1084/jem.20131241. URL: http://jem.rupress.org/content/210/12/2569.

Schmiedel, Benjamin J et al. (2018). "Impact of Genetic Polymorphisms on Human Immune Cell Gene Expression". *Cell* 175, pp. 1701–1715. DOI: 10.1016/j.cell.2018.10.022.

Schöfl, Gerhard et al. (2017). "2.7 million samples genotyped for HLA by next generation sequencing: lessons learned". *BMC Genomics* 18.1, pp. 1–16.

Schones, Dustin E et al. (2008). "Dynamic regulation of nucleosome positioning in the human genome". *Cell* 132.5, pp. 887–898.

Schraivogel, Daniel et al. (2020). "Targeted Perturb-seq enables genome-scale genetic screens in single cells". *Nature Methods* 17.6, pp. 629–635.

SEARCH Study Group et al. (2004). "SEARCH for Diabetes in Youth: a multicenter study of the prevalence, incidence and classification of diabetes mellitus in youth". *Controlled Clinical Trials* 25.5, pp. 458–471.

Shabalin, Andrey A (2012). "Matrix eQTL: ultra fast eQTL analysis via large matrix operations". *Bioinformatics* 28.10, pp. 1353–1358. DOI: 10.1093/bioinformatics/bts163.

Shapiro, AM James, Marta Pokrywczynska, and Camillo Ricordi (2017). "Clinical pancreatic islet transplantation". *Nature Reviews Endocrinology* 13.5, pp. 268–277.

Sharp, Seth A et al. (2019). "Development and standardization of an improved type 1 diabetes genetic risk score for use in newborn screening and incident diagnosis". *Diabetes Care* 42.2, pp. 200–207.

Singal, DP and MA Blajchman (1973). "Histocompatibility (HL-A) antigens, lymphocytotoxic antibodies and tissue antibodies in patients with diabetes mellitus". *Diabetes* 22.6, pp. 429–432.

Skyler, Jay S (2018). "Hope vs hype : where are we in type 1 diabetes?" *Diabetologia* 61, pp. 509–516.

Slatkin, Montgomery (2008). "Linkage disequilibrium—understanding the evolutionary past and mapping the medical future". *Nature Reviews Genetics* 9.6, pp. 477–485.

Small, Kerrin S et al. (2011). "Identification of an imprinted master trans regulator at the KLF14 locus related to multiple metabolic phenotypes". *Nature Genetics* 43.6, pp. 561–4.

Smith, Mia J. et al. (2015). "Loss of anergic B cells in prediabetic and new-onset type 1 diabetic patients". *Diabetes* 64 (5), pp. 1703–1712. ISSN: 1939327X. DOI: 10.2337/db13-1798.

Smith, P Jason et al. (2020). "PEPATAC: An optimized ATAC-seq pipeline with serial alignments". *bioRxiv*. DOI: https://doi.org/10.1101/2020.10.21.347054.

Smyth, Deborah J et al. (2006). "A genome-wide association study of nonsynonymous SNPs identifies a type 1 diabetes locus in the interferon-induced helicase (IFIH1) region". *Nature Genetics* 38.6, pp. 617–619.

Speed, Doug et al. (2012). "Improved heritability estimation from genome-wide SNPs". *American Journal of Human Genetics* 91.6, pp. 1011–1021.

Spielman, Richard S, Ralph E McGinnis, and Warren J Ewens (1993). "Transmission Test for Linkage Disequilibrium: The Insulin Gene Region and Insulin-dependent Diabetes Mellitus (IDDM)". *American Journal of Human Genetics* 52, pp. 506–516.

Steck, Andrea K. et al. (2011). "Age of islet autoantibody appearance and mean levels of insulin, but not GAD or IA-2 autoantibodies, predict age of diagnosis of type 1 diabetes: Diabetes autoimmunity study in the young". *Diabetes Care* 34 (6), pp. 1397–1399. ISSN: 01495992. DOI: 10.2337/dc10-2088.

Steck, Andrea K. et al. (2016). "Predictors of slow progression to diabetes in children with multiple islet autoantibodies". *Journal of Autoimmunity* 72, pp. 113–117. ISSN: 0896-8411. DOI: 10.1016/J.JAUT.2016.05.010. URL: http://www.sciencedirect.com/science/article/pii/S0896841116300713?via%3Dihub.

Steck, Andrea K et al. (2017a). "Can non-HLA single nucleotide polymorphisms help stratify risk in TrialNet relatives at risk for type 1 diabetes?" *The Journal of Clinical Endocrinology & Metabolism* 102.8, pp. 2873–2880.

Steck, Andrea K et al. (2017b). "Residual beta-cell function in diabetes children followed and diagnosed in the TEDDY study compared to community controls". *Pediatric Diabetes* 18.8, pp. 794–802.

Stone, Virginia M et al. (2018). "A Coxsackievirus B vaccine protects against virus-induced diabetes in an experimental mouse model of type 1 diabetes". *Diabetologia* 61.2, pp. 476–481.

Szklarczyk, Damian et al. (2017). "The STRING database in 2017 : quality-controlled protein – protein association networks, made broadly accessible". *Nucleic Acids Research* 45, pp. D362–D368. DOI: 10.1093/nar/gkw937.

Taliun, Daniel et al. (2021). "Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program." *Nature* 590.7845, pp. 290–299. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03205-y. URL: http://www.ncbi.nlm.nih.gov/pubmed/33568819.

Taub, M. A. et al. (2012). "Incorporating Genotype Uncertainties Into the Genotypic TDT for Main Effects and Gene-Environment Interactions". *Genetic Epidemiology* 36.3, pp. 225–234. DOI: 10.1002/gepi.21615.Incorporating.

The International HapMap Consortium (2003). "The international hapmap project". *Nature* 426.6968, pp. 789–796.

Thomas, Nicholas J et al. (2018). "Frequency and phenotype of type 1 diabetes in the first six decades of life: a cross-sectional, genetically stratified survival analysis from UK Biobank". *The Lancet Diabetes & Endocrinology* 6.2, pp. 122–129.

Thurman, Robert E et al. (2012). "The accessible chromatin landscape of the human genome". *Nature* 489.7414, pp. 75–82.

Todd, John A, John I Bell, and Hugh O McDevitt (1987). "HLA-DQ β gene contributes to susceptibility and resistance to insulin-dependent diabetes mellitus". *Nature* 329.6140, pp. 599–604.

Todd, John A. et al. (2007). "Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes." *Nature Genetics* 39.7, pp. 857–64. ISSN: 1061-4036. DOI: 10.1038/ng2068.

Traherne, J. A. et al. (2016). "KIR haplotypes are associated with late-onset type 1 diabetes in European-American families". *Genes and Immunity* 17 (1), pp. 8–12. ISSN: 14765470. DOI: 10.1038/gene.2015.44.

Trynka, Gosia et al. (2015). "Disentangling the Effects of Colocalizing Genomic Annotations to Functionally Prioritize Non-coding Variants within Complex-Trait Loci." *American Journal of Human Genetics* 97.1, pp. 139–52. ISSN: 1537-6605. DOI: 10.1016/j.ajhg.2015.05.016. URL: http://www.ncbi.nlm.nih.gov/pubmed/26140449http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4572568.

Tsukumo, Shin-ichi et al. (2013). "Bach2 maintains T cells in a naive state by suppressing effector memory-related genes". *PNAS* 110.26, pp. 10735–10740. DOI: 10.1073/pnas.1306691110.

Törn, Carina et al. (2015). "Role of type 1 diabetes- Associated snps on risk of autoantibody positivity in the TEDDY study". *Diabetes* 64 (5), pp. 1818–1829. ISSN: 1939327X. DOI: 10.2337/db14-1497.

Törn, Carina et al. (2016). "Complement gene variants in relation to autoantibodies to beta cell specific antigens and type 1 diabetes in the TEDDY Study". *Scientific Reports* 6 (1), p. 27887. ISSN: 2045-2322. DOI: 10.1038/srep27887. URL: http://www.nature.com/articles/srep27887.

Valdes, Ana M et al. (2012). "Use of class I and class II HLA loci for predicting age at onset of type 1 diabetes in multiple populations". *Diabetologia* 55.9, pp. 2394–2401.

Varshney, Arushi et al. (2017). "Genetic regulatory signatures underlying islet gene expression and type 2 diabetes". *PNAS* 114.9, pp. 2301–2306. DOI: 10.1073/pnas.1621192114.

Vella, Adrian et al. (2005). "Localization of a type 1 diabetes locus in the IL2RA/CD25 region by use of tag single-nucleotide polymorphisms". *American Journal of Human Genetics* 76.5, pp. 773–779.

Verlouw, Joost AM et al. (2021). "A comparison of genotyping arrays". *European Journal of Human Genetics*, pp. 1–14.

Voorter, Christina EM, Fausto Palusci, and Marcel GJ Tilanus (2014). "Sequence-based typing of HLA: an improved group-specific full-length gene sequencing approach". *Bone Marrow and Stem Cell Transplantation*. Springer, pp. 101–114.

Võsa, Urmo et al. (2018). "Unraveling the polygenic architecture of complex traits using blood eQTL meta- analysis". *bioRxiv*, pp. 1–57.

Wallace, Chris et al. (2015). "Dissection of a Complex Disease Susceptibility Region Using a Bayesian Stochastic Search Approach to Fine Mapping". *PLoS Genetics* 11.6. Ed. by Jonathan Marchini, e1005272. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1005272.

Wand, Hannah et al. (2021). "Improving reporting standards for polygenic scores in risk prediction studies". *Nature* 591.7849, pp. 211–219.

Wang, Gao et al. (2020). "A simple new approach to variable selection in regression, with application to genetic fine mapping". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 82.5, pp. 1273–1300.

Wang, Kai, Mingyao Li, and Hakon Hakonarson (2010). "ANNOVAR : functional annotation of genetic variants from high-throughput sequencing data". *Nucleic Acids Research* 38.16, e164. DOI: 10.1093/nar/gkq603.

Wang, Zhong, Mark Gerstein, and Michael Snyder (2009). "RNA-Seq: a revolutionary tool for transcriptomics". *Nature Reviews Genetics* 10.1, pp. 57–63.

Ward, Lucas D and Manolis Kellis (2016). "HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease". *Nucleic Acids Research* 44, pp. 877–881. DOI: 10.1093/nar/gkv1340.

Wellcome Trust Case Control Consortium (2007). "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls." *Nature* 447.7145, pp. 661–78. ISSN: 1476-4687. DOI: 10.1038/nature05911. URL: http://www.ncbi.nlm.nih.gov/pubmed/17554300http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2719288.

Westra, Harm-jan et al. (2013). "Systematic identification of trans eQTLs as putative drivers of known disease associations". *Nature Genetics* 45.10, pp. 1238–1243. ISSN: 1061-4036. DOI: 10.1038/ng.2756.

Westra, Harm Jan et al. (2018). "Fine-mapping and functional studies highlight potential causal variants for rheumatoid arthritis and type 1 diabetes". *Nature Genetics* 50.10, pp. 1366–1374. ISSN: 15461718. DOI: 10.1038/s41588-018-0216-7.

Wickham, Hadley (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN: 978-3-319-24277-4. URL: https://ggplot2.tidyverse.org.

Wilke, Claus O. (2020). *cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'*. R package version 1.1.1. URL: https://CRAN.R-project.org/package=cowplot.

Willer, Cristen J., Yun Li, and Goncalo R. Abecasis (2010). "METAL: Fast and efficient meta-analysis of genomewide association scans". *Bioinformatics* 26.17, pp. 2190–2191. ISSN: 13674803. DOI: 10.1093/bioinformatics/btq340.

Wing, Kajsa et al. (2008). "CTLA-4 control over Foxp3+ regulatory T cell function." *Science* 322 (5899), pp. 271–5. ISSN: 1095-9203. DOI: 10.1126/science.1160062. URL: http://www.ncbi.nlm.nih.gov/pubmed/18845758.

Wojcik, Genevieve L et al. (2019a). "Genetic analyses of diverse populations improves discovery for complex traits." *Nature* 570.7762, pp. 514–518. ISSN: 1476-4687. DOI: 10.1038/s41586-019-1310-4. URL: http://www.ncbi.nlm.nih.gov/pubmed/31217584http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6785182.

Wojcik, GL et al. (2019b). "The future is now: genomic studies must be globally representative". *European Journal of Human Genetics* 27, pp. 1111–1111.

Wollenhaupt, Jürgen et al. (2019). "Safety and efficacy of tofacitinib for up to 9.5 years in the treatment of rheumatoid arthritis: final results of a global, open-label, long-term extension study". *Arthritis Research & Therapy* 21.1, p. 89.

WTCCC2 Studies. https://www.wtccc.org.uk/ccc2/wtccc2_studies.html. Accessed: 2021-09-02.

Yang, Jian et al. (2011). "GCTA: a tool for genome-wide complex trait analysis". *American Journal of Human Genetics* 88.1, pp. 76–82.

Yin, Tengfei, Dianne Cook, and Michael Lawrence (2012). "ggbio: an R package for extending the grammar of graphics for genomic data". *Genome Biology* 13.8, pp. 1–14.

Yoshihara, Eiji et al. (2020). "Immune-evasive human islet-like organoids ameliorate diabetes". *Nature* 586.7830, pp. 606–611.

Zhao, Lue Ping et al. (2016). "Next-Generation Sequencing Reveals That HLA-DRB3, -DRB4, and -DRB5 May Be Associated With Islet Autoantibodies and Risk for Childhood Type 1 Diabetes." *Diabetes* 65 (3), pp. 710–8. ISSN: 1939-327X. DOI: 10.2337/db15-1115. URL: http://www.ncbi.nlm.nih.gov/pubmed/26740600http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4764147.

Zhao, Lue Ping et al. (2017). "Building and validating a prediction model for paediatric type 1 diabetes risk using next generation targeted sequencing of class II HLA genes". *Diabetes/Metabolism Research and Reviews* 33.8, e2921.

Zhu, Meng et al. (2019). "Identification of Novel T1D Risk Loci and Their Association With Age and Islet Function at Diagnosis in Autoantibody-Positive T1D Individuals: Based on a Two-Stage Genome-Wide Association Study." *Diabetes Care* 42.8, pp. 1414–1421. ISSN: 1935-5548. DOI: 10.2337/dc18-2023. URL: http://www.ncbi.nlm.nih.gov/pubmed/31152121.

Ziegler, Anette G et al. (2013). "Seroconversion to multiple islet autoantibodies and risk of progression to diabetes in children". *JAMA* 309.23, pp. 2473–2479.