Integrating empirical data and population genetic simulations to study the genetic architecture of type 2 diabetes

Integrating empirical data and population genetic simulations to study the genetic architecture of type 2 diabetes

---

A dissertation presented

by

Vineeta Agarwala

to

The Harvard University Program in Biophysics

in partial fulfillment of the requirements
for the degree of

Doctor of Philosophy

in the subject of

Biophysics

Harvard University
Cambridge, Massachusetts
August 2013

Dissertation Advisor: Professor David M. Altshuler Vineeta Agarwala

Integrating empirical data and population genetic simulations to study the genetic architecture of type 2 diabetes

Abstract

Most common diseases have substantial heritable components but are characterized by complex inheritance patterns implicating numerous genetic and environmental factors. A longstanding goal of human genetics research is to delineate the *genetic architecture* of these traits – the number, frequencies, and effect sizes of disease-causing alleles – to inform mapping studies, elucidate mechanisms of disease, and guide development of targeted clinical therapies and diagnostics. Although vast empirical genetic data has now been collected for common diseases, different and contradictory hypotheses have been advocated about features of genetic architecture (e.g., the contribution of rare vs. common variants). Here, we present a framework which combines multiple empirical datasets and simulation studies to enable systematic testing of hypotheses about both global and locus-specific complex trait architecture. We apply this to type 2 diabetes (T2D).

For T2D, we find that extreme models of global genetic architecture are excluded (e.g., models where T2D is a collection of rare Mendelian diseases), but a wide range of models remain consistent with epidemiology, linkage, and genome-wide association studies (GWAS). Simulations predict that ongoing sequencing and genotyping studies (in tens of thousands of individuals) will further constrain architecture, but that very large sample sizes (e.g., >250K unselected individuals) will be required to localize most T2D heritability.

To characterize allelic architecture at individual T2D loci, we develop haplotype-based methods to integrate data from GWAS and low-pass sequencing of thousands of T2D cases and controls. We find varied architectures plausible at each locus. At some loci, the most likely model implicates common causal variation (chr9p21, TCF7L2, KCNJ11, HNF1B). At others, there is evidence for common variants of weak effect alongside independent low-frequency variants of larger

effect (CCND2, KCNQ1) or a burden of very rare protein-coding, disease-associated mutations (PPARG). Finally, at several loci, further genetic and/or experimental interrogation is required to determine whether causal alleles are common, rare, or both (HMGA2, IGF2BP2).

In this thesis, we have integrated diverse datasets to better understand the genetic and biological architecture of T2D. This work informs future genetic and experimental studies of T2D, and provides methods for hypothesis testing that are broadly applicable to many complex traits.

Table of Contents

Acknowledgements

The work presented in this thesis would not have been possible without the incredible support, mentorship, and friendship of so many kind individuals.

Foremost, I owe a huge debt of gratitude to David Altshuler. I thank my lucky stars for having met David at one of the first MD/PhD program 'Meet the Investigator' Lunch seminars during the summer of 2008, before graduate school had even begun. I remember him describing his own career path and dogged determination to study the genetics of common human diseases, and I remember recognizing in that moment that this was a mentor who I knew would inspire me throughout my PhD training. Since the time I joined the lab (in January of 2009), David has been an even greater mentor, teacher, and life guide than I could have ever imagined. He begins one-on-one meetings by asking the question "Are you having fun?", ends them with "This is going to be amazing", and sprinkles so much advice along the way that members of the lab often joke that tape recording our meetings with David might be good not only for our science but also for our general morale. I have learned much from David about the process of scientific research, about the current and future state of human genetics as a field, about writing (in such a way that our work can be impactful for broad audiences), and about making important personal and professional life decisions. I am so grateful for the opportunities David has made possible for me, and for the confidence he placed in me along the way.

This thesis would simply never have been completed in its current form without the tremendous mentorship of Jason Flannick (a post-doctoral fellow in David's lab). Jason not only conceived of the original idea to create a population genetic simulation framework for complex disease, but he also played an instrumental role (with his computer science savvy) in bringing me into the light – that is, in teaching me to use a pipeline software he developed that not only advanced my own graduate work but may also soon change the way the Broad analyzes large-scale biological data of many kinds. Beyond this, Jason taught me – through conversations we have had almost

every single day during my PhD – about how to frame a hypothesis, how to ask a precise question, how to answer these questions with meticulous analysis, how to present data, and how to lead amongst a group of collaborators. Jason is that rare scientist who can (during the same conversation) enlighten you about the fastest, most amazing UNIX command you have ever seen, point you to some clutch advice from Strunk and White, and also be the harshest critic of the most recent genetic analysis you did. No matter how busy he is, Jason will put down what he is doing to think through a scientific problem or offer advice and I am so grateful that I had the opportunity to work so closely with Jason on a number of projects over the past few years.

While working at the Broad Institute and MIT, I have also been fortunate to be the recipient of both scientific and professional guidance from so many incredible scientists. I would like to thank my thesis committee members – Leonid Mirny, Soumya Raychaudhuri, and Manolis Kellis – for treating me as if I was a member of their own labs and so generously sharing their time with me. I was thrilled to have Leonid as both a professor as well as a research advisor (on the publication described in Appendix A3). Soumya gave me countless hours of career and research advice for which I am so grateful (and he recently gave me the cool opportunity to present at his lab meeting, which was great fun). Manolis has been a close collaborator and advisor on a number of projects, and I have appreciated his making me always feel as if I was part of his extended 'lab family'. I could not have asked for a more supportive, inquisitive, and constructive thesis committee. Outside my thesis committee, I owe huge thanks to so many others who have taken the time to review my work, give me feedback, or just chat about science: Joel Hirschhorn (who also kindly chaired my thesis examination committee), Mark Daly, Shamil Sunyaev, Eric Lander, Steve McCarroll, Mark DePristo, Ben Voight, Ben Neale, Steve Schaffner, and Feng Zhang have been just a few of these special people. Thank you for taking the time to guide a young scientist!

I am thankful to a number of other mentors who have helped me chart a path through the first two years of medical school and PhD training as part of the Biophysics graduate program. Thank you, David Cohen, Rick Mitchell, and Patty Cunningham, for all you have done for me already

during the first phase of this dual-degree program. Thank you, Jim Hogle and Michele Jakoulov, for your tireless efforts in keeping the Biophysics family healthy and happy all year long. Finally, thank you, Anne Giersch and David Housman, for giving me the tremendous opportunity to serve as a teaching assistant in the HST human genetics course for three years! That was such an honor, and such a fun way to discover that I love teaching.

The Altshuler-Daly lab is a really remarkable group, and it has been so much fun working alongside friends in this lab. When I first joined the lab, Jessica Alston and Anthony Philippakis were gracious enough to accept me as their 'roton', and took the time to teach me about everything from genotype file formats to SNP calling algorithms to cell culture and enhancer assays. I learned so much from these two individuals, and they were instrumental in confirming my (naïve at the time) interest in human genetics. Jessica was then in the lab for several years after I joined; thank goodness for her friendship, and for all the 'across-the-wall' conversations we had! Amit Majithia has been my desk-mate for the past three years, and this was sheer coincidence that I was paired with someone with whom I went on to connect so well. Every single morning, Amit has been a human barometer of sorts, sensing whether there is something on my mind or something I need to talk about; thank you, Amit, for being a general sage, a scientific (and clinical) role model, and a fabulous friend. Chris Hartl, my crazy smart friend who 'casually' keeps an in-depth online blog with solutions to hard problems in human genetics: thank you for reminding me to read, for your willingness to do things like create 'DumbSim' on a random evening, and for listening. Jason Wright, god of non-coding functional biology: thank you for encouraging me to think big, for discussing your brilliant ideas with me, and for being such a wonderful partner in crime on our chr9p21 studies!

Graduate school would not have been the same without the support network of so many lovely friends. One of my closest friends – also a scientist – often asked me at the end of a long day: "So what did you learn today?" It was questions like these that kept me grounded – and made me grateful for the chance to work in a field where a day when I *didn't* learn something was rare! I feel so fortunate to have such special friends surround me all through these past few years, and it

has so been so much fun to be a part of the big moments in their lives too. Thank you to my medical school classmates, especially Wen, Yian, and Sonali. Thank you to my grad school and teaching buddies: Lizzy (also lab-mate and conference traveling partner extraordinaire), Jesse, Chewie, and Mark. Thank you, above all, to all the caring friends who kept me sane along the way, especially Naomi, Priyanka, Shivan, and Vikram. You are all simply the best.

Finally, no words can even begin to express the appreciation that I feel for my family: my mom and dad, and my siblings, Vivek and Vandana. My parents have been the source of tremendous personal and professional inspiration: their work ethic, commitment to getting it right, and enthusiasm for their day-to-day roles has always set an incredible example for me. I am so grateful to my siblings for slogging through so many long car rides to visit me in Boston, for taking the time to talk to me every day, and for being fun-loving and reliable rocks in my life. I love you guys. I am blessed to have you as my family, and to have had your support during these past years.

Motivation and thesis overview

The central goal of this thesis is to systematically test hypotheses about the global and locus-specific genetic architecture of the complex human disease type 2 diabetes (T2D).

Why study type 2 diabetes (T2D)? T2D is one of the most common human genetic diseases (affecting over 8% of the U.S. population), and is a leading cause of morbidity (causing blindness, kidney disease, nervous system damage, amputation) and mortality (a significant risk factor for heart disease and stroke).[1] Existing treatments for T2D are of limited efficacy, and no available drugs can reverse or even halt progression of the disease. The development of new therapies for T2D will require deeper understanding of its pathogenesis in human populations. Human genetic studies provide a powerful and unbiased window into the genes, pathways, and processes that causally contribute to the onset of disease.

Why study the genetic architecture of complex traits like T2D? The genetic architecture of a disease describes the answers to many questions of interest. How many genetic mutations across the genome contribute to risk of the disease? How many mutations does *each individual patient* carry? How frequent are these mutations in the population? Do individuals have their own private mutations, or is the majority of disease attributable to mutations common across the population? And finally, by how much does each mutation increase or decrease risk of disease? Do a very large number of mutations each modify disease risk very slightly, or do some mutations make an individual much more or less likely to be afflicted with disease?

These questions have profound implications in both clinical and research realms. The extent to which 'personalized' clinical medicine will ever be possible for a disease like T2D, for example, depends on the underlying spectrum of disease-causing genetic variation: targeted diagnosis and treatment based on individual genome sequence will be more tractable if the disease is caused by rare mutations of large effect than if many genes and variants together contribute.[2–5] The genetic

architecture of a trait also governs the success of future genetic mapping studies[6–9], and informs on the choice of experimental designs and analytical methods that are optimally powered to uncover novel signals[10,11]. Thus, understanding the genetic architecture of a trait is an important objective, both in order to design efficient future research studies and also to guide the translation of genetic information to clinical settings.

A number of genetic studies have been performed in recent years to probe the genetic basis of complex human diseases, but it is challenging to quantify the constraints these data place on the underlying genetic architecture. In this thesis, we address this question – focusing on T2D – in several ways. First (in Chapters 1-3) we develop a simulation-based framework that is calibrated to empirical data and enables systematic testing of hypotheses about genetic architecture. We begin by using forward evolutionary simulation to model human genetic variation in hundreds of thousands of individuals, recapitulating properties of empirical sequencing data (Chapter 1). We then develop a principled set of simple, population genetic parameters to control the mapping of genotype to phenotype, which results in a wide space of potential genetic architectures for T2D (Chapter 2). To evaluate each of these architectures, we perform (*in silico*) a number of different genetic studies, as they were conducted for T2D, under each simulated model and ask which models produce results consistent with an array of empirical observations (Chapter 3). This work defines a novel approach in which many hypotheses about the architecture of a given trait can be simultaneously tested against an integrated panel of empirical data for the trait.

In the next major section of this thesis (Chapters 4-6), we move from exploring the *global* genetic architecture of a disease to studying the *local* allelic architecture, or the spectrum of causal allele frequencies and effect sizes at individual disease loci across the genome. In Chapter 4, we simulate the diverse panel of genetic architectures developed in Chapter 1-2 within human genes, generating thousands of simulated loci (in thousands of samples) with varying numbers of causal variants, each with different frequencies and effect sizes. We use these simulations to evaluate the implications of genetic architecture on the power (and therefore choice) of different analytical

methods. Specifically, we focus on the class of gene-based rare variant association methods; a large number of such methods have been published in recent years, yet they have not been systematically compared and evaluated under a broad range of locus architectures. We describe the power of each method (at varying significance thresholds), and learn which methods are best powered to test different hypotheses about complex disease.

In Chapters 5-6, we dive into characterizing the allelic architecture empirically observed at loci identified by GWAS for T2D (loci where common variants are known to be robustly associated with risk of T2D). Armed with the insights learned in simulation studies from the first half of this thesis, we integrate GWAS and whole genome sequence data in unrelated cases and controls to test three principle hypotheses at these loci: (1) that common variant(s) causally modulate risk of T2D; (2) that rare variants create 'synthetic' common variant associations; and (3) that rare variants (individually or in aggregate) have effects on T2D, independent of the common signals. We characterize the architecture at a single, fascinating locus (chr9p21) in great depth (Chapter 5), and use this locus as a testing ground to compare different genotyping vs. imputation-based fine-mapping strategies as well as develop a set of new haplotype-based methods for use with genome sequencing data. We then apply these methods to ten other T2D-associated loci (Chapter 6), demonstrating that different genetic models are plausible at each locus. In each case, we enumerate sets of candidate causal variants (common and rare) for use in functional follow-up studies of these loci.

Finally, recognizing that human genetic studies are only the first step on the path towards actually gaining insight into the pathophysiology of a disease like T2D, we next attempt to integrate some of this genetic data with *experimental* biological datasets. Because the vast majority of T2D loci identified in GWAS localize to "non-coding" regions of the genome (regions that do not contain protein-coding exons), we focus in these chapters on characterizing the function of putative regulatory elements – regions which do not themselves encode proteins, but may function to control and modulate the expression of nearby genes. Specifically, we perform *in vitro* tiling screens for

regulatory function across two T2D loci (chr9p21 and JAZF1; Chapter 7), and identify promising enhancers for further follow-up. We also look for signatures of genome-wide enrichment for variants associated to T2D across elements predicted to have regulatory activity (Chapter 8); this analysis reveals broad enrichment for T2D association across regulatory elements in many cell types including adipocytes, hepatocytes, and pancreatic islets.

References

1. National Center for Chronic Disease Prevention and Health Promotion National Diabetes Fact Sheet, 2011. (2011).
2. Collins, F.S. & McKusick, V. Implications of the Human Genome Project for Medical Science. *JAMA* **285**, 540-4 (2001).
3. Jostins, L. & Barrett, J.C. Genetic risk prediction in complex disease. *Human Molecular Genetics* **20**, R182-8 (2011).
4. Thanassoulis, G. & Vasan, R. Genetic Cardiovascular Risk Prediction - Will We Get There? *Circulation* **122**, 2323-2334 (2011).
5. Grant, R.W., Moore, A.F. & Florez, J.C. Genetic architecture of type 2 diabetes: recent progress and clinical implications. *Diabetes Care* **32**, 1107-14 (2009).
6. Altshuler, D., Daly, M.J. & Lander, E.S. Genetic mapping in human disease. *Science* **322**, 881-8 (2008).
7. Hirschhorn, J.N. & Daly, M.J. Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics* **6**, 95-108 (2005).
8. McCarthy, M.I. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics* **9**, 356-69 (2008).
9. Cirulli, E.T. & Goldstein, D.B. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature Reviews Genetics* **11**, 415-25 (2010).
10. Kryukov, G.V., Shpunt, A., Stamatoyannopoulos, J.A. & Sunyaev, S.R. Power of deep, all-exon resequencing for discovery of human trait genes. *PNAS* **106**, 3871-6 (2009).
11. Price, A.L. *et al.* Pooled association tests for rare variants in exon-resequencing studies. *American Journal of Human Genetics* **86**, 832-8 (2010).

Background

Human disease phenotypes have an inherited component

It has long been appreciated that traits ('phenotypes') which we observe in the natural world are inherited in successive generations. In the domestication of animals and plants, it has been known for centuries that selective breeding can enrich for desirable characteristics. In the study of human phenotypes, it was also clear that children resembled their parents. As William Bateson (who also gave the field of 'genetics' its name) wrote in 1908, "That we are assemblages or medleys of our parental characteristics is obvious. We all know that a man may have his father's hair, his mother's color, his father's voice, his mother's insensibility to music, and so on."[1]

Two key discoveries in the late nineteenth century – made around the same time, but independently of one another – laid the foundation for modern understanding of genetics. In 1859, Charles Darwin published *The Origin of Species*, describing his theory of natural selection and evolution. In order for this theory to work, individuals must exhibit phenotypic variability, and must further be able to pass on these phenotypes to offspring; Darwin did not, however, know the mechanism of this transmission. In 1865, Gregor Mendel published *Experiments in Plant Hybridization*[2], in which he observed a "striking regularity with which the same hybrid forms always" appeared among offspring arising from artificial fertilization of pea plants. These experiments – though not appreciated until decades later – established the mathematical rules of dominant and recessive inheritance, and suggested that (at least for some traits) inheritance occurs via the independent transmission of discrete units.

By the turn of the century, the inheritance patterns of some human traits were recognized to bear resemblance to the rules Mendel described. In 1902, Archibald Garrod observed higher incidence of the disease alkaptonuria ('black urine disease') among offspring of consanguineous parents, and reported that "the law of heredity discovered by Mendel offers a reasonable account of such [recessive] phenomena."[3] Though the exact gene defect underlying alkaptonuria was not

discovered until the 1960s, this trait became among the first-described human diseases recognized to be caused by a single gene defect (and transmitted in a Mendelian fashion). By 1925, shortly after Garrod's report, the inheritance of human blood groups was also shown to be explained by a system of triple alleles that were passed onto offspring in a manner consistent with Mendel's laws.

In parallel to these observations about inheritance patterns, biologists were also forming theories about the physical matter by which heredity could operate. In the mid-1880s, chromosomes were identified under the microscope (Boveri), and were recognized to occur in pairs (one maternal copy, and one paternal copy). By 1903, W. Sutton had documented that pairs of chromosomes oriented at random on meiotic spindles, thus raising the "probability that the association of paternal and maternal chromosomes in pairs and their subsequent separation during the reducing division…may constitute the physical basis of the Mendelian law of heredity."[4]

The confluence of chromosomal theory and Mendel's inheritance laws ultimately resulted in the recognition of the property of *genetic linkage*. As early as 1900, it was observed by Carl Correns that some traits are more likely to be inherited *together* rather than independently (as Mendel's laws would have predicted). In 1910, Thomas Hunt Morgan reported the sex-linked inheritance of white eyes (and other traits) in Drosophila, suggesting that the genes underlying these traits were physically coupled to the genes determining sex (e.g., on the X chromosome). The idea of "linkage groups" was developed to refer to the idea that genes on the same chromosomes were more likely to be inherited together.

It was also realized, however, that recombination ("crossing over") could occur between such groups, and that the likelihood of recombination depended on the distance between two genes. In 1913, Alfred Sturtevant (then a student in Morgan's laboratory), developed the first "linkage map" of a chromosome, using the strength of linkage (measured by co-inheritance) between genes on the Drosophila sex chromosome to deduce their approximate physical distance from one another.[5] This concept of linkage would go on to have profound effects on the field of human genetics.

In order for the impact of such linkage maps to be realized, one key piece of the puzzle was needed: an understanding of genetic *mutation*. How does variability in genes arise in the first place? In Drosophila studies, novel versions (*alleles*) of the same gene were recognized when new phenotypes spontaneously appeared in fly lineages. Darwin's original work, too, had assumed the presence of "fluctuating variations" as the substrate on which natural selection acts.[4] In order for these hypothetical concepts of mutation to become concrete, however, the discovery of DNA (by James Watson and Francis Crick in 1953), followed by the development of methods to read out DNA sequence at sites in the genome (by Frederick Sanger, Walter Gilbert, and Allan Maxam in the 1970s), was required. Following this, sites of naturally occurring DNA polymorphisms (locations at which individuals commonly have different alleles) were identified across the human genome. These could then be used to construct *human* linkage maps (as Sturtevant had made for Drosophila) and trace the transmission of chromosomal regions through families.[6]

By systematically correlating disease status with the transmission of particular alleles at polymorphic markers across the genome, it became possible to identify marker sites (and chromosomal regions) with which putative disease-causing alleles must be *linked*. This advent of *genetic mapping* in humans resulted in the localization of genes underlying hundreds of 'Mendelian' disease phenotypes, ranging from Huntington Disease (in 1983) to cystic fibrosis (in 1988).[7]

Common human diseases are 'complex', not Mendelian

The mapping of Mendelian human traits has been hugely important: it has enabled diagnosis and (in some cases) treatment for a range of severe diseases, and further propelled forward our understanding of the biological processes contributing to disease pathophysiology. However, most common human diseases (indeed, most phenotypes in general) actually *do not* show Mendelian patterns of inheritance. Mendelian inheritance requires several features: a single genetic defect must be *sufficient* to produce the disease (thus, a single mutation must have an effect on phenotype), and it must also be *necessary* (there must not be other genetic or non-genetic causes of disease). The diseases that affect most people around us – type 2 diabetes (T2D), hypertension,

heart disease, mental illness, and others – clearly have an inherited basis, but do not obey these Mendelian properties, and do not show patterns of recessive or dominant transmission in families.

As a simple example of such 'complexity' in human phenotypes, Sturtevant reported (in 1940) on human inheritance of the trait tongue-rolling. He compared the ability of parents and their offspring to roll their tongue (**Figure B1**), and observed confusing results: rolling appeared to be partially dominant, but two non-rolling parents could nonetheless have children capable of rolling their tongue. This suggested that multiple genes might influence this trait, or perhaps that environmental influences were substantial (e.g. the trait could be learned). In subsequent years, as more families (and twins) were studied, the picture grew more confusing and Sturtevant later wrote that "he was embarrassed to see it listed…as an established Mendelian case."[4]

Parents	R offspring	NR offspring
R x R	28	5
R x NR	33	22
NR x NR	4	9

Figure B1: Sturtevant's results on inheritance of tongue-rolling

Such is the story for most common human phenotypes – their inheritance patterns appear complex. Studies of human height provide the classical example of complex inheritance. As early as 1886, Francis Galton invented the concept of regression while studying the height of parents and their offspring.[8] He found that while parents' height clearly influenced the height of offspring, the relationship was not deterministic, and only a fraction of the variance in offspring height could be explained by the parents' phenotypes. This fraction (related to the slope of the regression) came to be known as the *heritability* of a complex trait. Galton coined the phrase "nature vs. nature" to refer to the substantial non-genetic (environmental) influences he observed on complex human phenotypes.

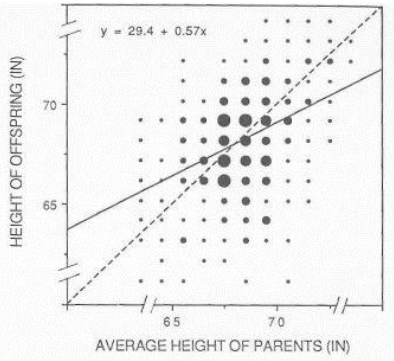


Figure B2: Galton's regression of offspring vs. mid-parent height

Galton's work, alongside that of others including Karl Pearson, gave rise to biometrics – a field concerned with the study of continuously varying traits (such as height), rather than traits showing discontinuous Mendelian inheritance. In 1918, R. A. Fisher published his seminal paper

The Correlation between Relatives on the Supposition of Mendelian Inheritance[9], in which he described novel statistical methods for analysis of variance (which are of course widely used today in many areas outside of genetics). In this work, Fisher reconciled the divide between biometricians and Mendelian geneticists, demonstrating that continuous phenotypes could result from the aggregate additive effects of many genetic factors, each of which could be inherited in a Mendelian fashion and individually produce only a small effect on the total phenotype (a 'polygenic' model).

Common human disease phenotypes such as type 2 diabetes, however, are not observed as continuous traits; rather, they are dichotomous. Evidence that such traits are heritable comes from the observation that the incidence of disease is higher among relatives (e.g. siblings) of affected individuals than it is in the general population. The sibling relative risk for T2D, for example, is ~2.0-2.5; this risk is lower than expected if T2D were inherited in a monogenic, Mendelian fashion, but the fact that the risk exceeds one suggests a genetic component (controlling for shared environment). In 1965, D. S. Falconer suggested that dichotomous traits might be studied *as if* a continuously varying trait was underlying them; disease could be thought to result above a threshold on this continuous 'liability' scale.[10,11] Falconer pioneered heritability estimation for dichotomous diseases by using data on incidence rates among relatives; in the next decades, narrow-sense heritability (the component of phenotypic variance attributable to additive genetic effects) was estimated for a host of human disease phenotypes and found, in most cases, to range from ~30-80%.[12,13]

Linkage mapping works for Mendelian traits, but fails for complex traits

Given that common disease traits have a significant inherited basis, geneticists were keen to apply linkage mapping – which had worked so well in the case of rare Mendelian disease phenotypes – to localize the causal genetic variation underlying these traits. This effort resulted in the identification of Mendelian sub-types of common diseases (e.g. familial forms of breast cancer caused by BRCA1/2 mutations, diabetes caused by mutations in one of several 'MODY' genes, or Alzheimer's caused by APP mutations), but only explained a small fraction (usually <1%) of the total incidence of these diseases. Moreover, these sub-types had different properties as compared to the

common forms of disease (they were usually more severe, exhibited earlier age of onset, and were not subject to environmental risk factors), making them distinct biological entities. Linkage mapping studies in large families or siblings with the *common* forms of these diseases yielded largely equivocal results.[7]

This finding was consistent with the biometric hypothesis that common diseases may be polygenic; that is, they may be caused by a large number of genetic mutations, such that no individual mutation (and no marker linked to it) shows any significant correlation with disease status. But it may also have simply been the expected result for traits with significant non-genetic components. This question is addressed in Chapter 3 of this thesis.

Genome-wide association studies identify numerous complex trait loci

In the wake of these negative linkage findings, population genetic theory offered a new path forward. Instead of tracing the transmission of disease mutations through families (where high penetrance and large effect sizes are required to observe an effect), what if the frequencies of millions of common polymorphisms across the genome could be compared between large, unrelated groups of affected and unaffected individuals?[7] The justification for such *genome-wide association studies* (GWAS), was termed the 'common disease common variant' (CDCV) hypothesis. This hypothesis was multi-factorial, but was ultimately grounded in population genetic assumptions about a) human demographic history and b) natural selection.[14,15] The human population was known to have grown exponentially after a bottleneck; it was therefore proposed that some deleterious alleles may have risen to common frequencies. It was further reasoned that these disease alleles might have been subject to only mild natural selection because disease phenotypes that are *common* likely have limited effects on reproductive fitness. Taken together, these simple assumptions suggested that perhaps systematic assay of common sites of variation could point to disease-relevant loci.

GWAS were enabled by rapid advances in genomic technology in the early 2000s.[16] After the human genome sequence was completed 2001, millions of common single-nucleotide

polymorphisms (positions at which two different alleles are common in the population) across the genome were systematically identified (by the International SNP Consortium). Spurred by rapid strides in the development of efficient and cost-effective SNP genotyping technologies, the International HapMap Project was next initiated; by 2007, this project had completed genotyping of over three million SNPs in 270 individuals from four ethnically diverse populations.[17,18] These data revealed that 'recombination hotspots' were scattered across the genome, and essentially provided granular, detailed descriptions of T. H. Morgan's 'linkage groups'. These linkage blocks (regions between recombination hotspots) exhibited low haplotype diversity and high correlation between the genotypes of neighboring SNPs. It was recognized that these groups could be exploited to perform GWAS efficiently: only one of a set of correlated SNPs must be queried, with little loss in information.

Over the next few years, genome-wide genotyping arrays containing ~500K common 'tags' for each unique haplotype within linkage blocks were developed, and the first large-scale GWAS were published in 2007 for a range of common human disease traits, including type 2 diabetes.[19–22] By 2012, over 1300 GWAS had been conducted for common diseases ranging from inflammatory bowel disease to coronary artery disease to schizophrenia to asthma. Because each GWAS tests hundreds of thousands of hypotheses, stringent standards for independent replication and statistical significance were established (after Bonferroni correction, only variants having an association p-value < 5*10^-8 were considered 'genome-wide significant'). In order to attain high power in such a setting, GWAS were conducted in very large sample sizes (tens of thousands of unrelated individuals, often in a staged design). A host of analytical methods were developed to correct for potential confounders such as population stratification.

GWAS successfully identified hundreds of haplotype blocks across the genome in which common markers show robust (and replicable) association to a variety of human diseases. In many cases, the association signals point to novel biology or pathways previously unconnected to disease. For example, GWAS of type 2 diabetes have identified over 70 genomic loci. Reassuringly, some of these GWAS loci (e.g. HNF4A) were previously known to harbor mutations that cause Mendelian

forms of early-onset diabetes. Several loci (e.g. KCNJ11) implicate pancreatic beta cell biology; given that T2D was thought to be principally a disease of insulin resistance in peripheral tissues, this was perhaps surprising. Finally, other loci (e.g. associations near the cell cycle regulators CDKN2A/B and CCND2; **Figure B3**) suggest biological mechanisms of disease that are, as of yet, entirely unknown.

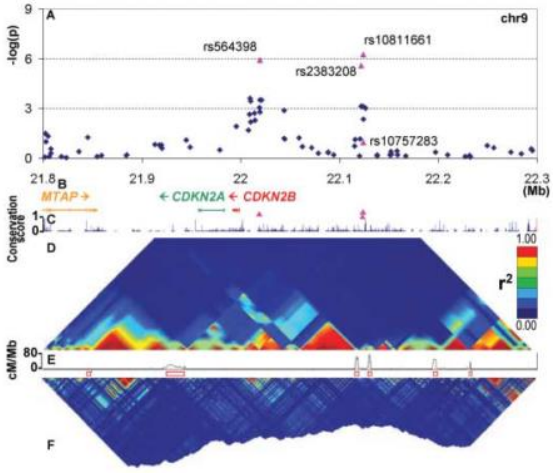


Figure B3: Example GWAS locus identified for T2D (at chr9p21, near CDKN2A/B).[58]

The translation of GWAS findings to actionable therapeutic and diagnostic insights, however, has been challenging. This has occurred for several reasons: (a) the associated markers are, in most cases, *not* directly disease-causing but rather just proxies for causal variation, (b) the linkage blocks implicated in GWAS are large, often spanning multiple different genes, (c) the associated variants often localize to poorly understood non-protein-coding regions of the genome, and (d) the effect sizes of disease-associated markers are, on average, very small (common marker odds ratios are typically less than 1.5, meaning that they increase disease risk by only a small amount). As a result, the therapeutic targets suggested by most GWAS loci are far from clear, let alone the direction in which these targets should be modified. The small effect sizes have made diagnosis based on genotype challenging; no individual variant is a strong predictor of disease, and even when taken together the GWAS loci do not (for most common diseases) improve risk prediction above what was previously achievable using family history and clinical factors.[23,24]

There are widely varied hypotheses about 'missing heritability' and genetic architecture

The total fraction of heritability explained by all the genome-wide significant loci discovered in GWAS has been limited for most common diseases (~10% for type 2 diabetes).[25] This so-called "missing heritability" of disease[26] has brought the debate about *genetic architecture* back to center

stage.[14,27–49] How many causal variants contribute to the risk of complex diseases, and what are their frequencies and effect sizes?

Even pre-GWAS, the idea that allelic heterogeneity (multiple causal alleles at a locus) would substantially reduce the power of common variant association scans had been appreciated[49]; some have argued that many loci have been undiscovered by GWAS for this reason. It has also been proposed that rare alleles (not tested by GWAS) could have larger effects than those detected at common GWAS alleles; disease mutations of large effect might have been kept rare due to purifying selection. In this case, rare variants (both at GWAS loci and elsewhere in the genome) might explain a large portion of yet-undiscovered genetic risk. At the extreme, some have even argued that common diseases are merely a poorly phenotyped aggregation of hundreds of Mendelian sub-types, suggesting that there may exist a very large number of individually rare, highly penetrant causal alleles.[29,31,35] The many rare variants revealed by recent exome sequencing studies[50–54] have been interpreted as supporting evidence for such hypotheses.

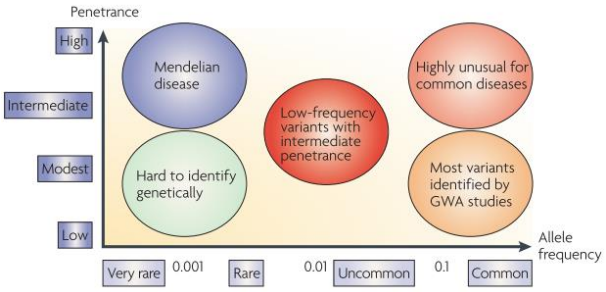


Figure B4: Potential role of variants of different frequency and effect size in complex trait genetic architecture.[33]

Another rare variant model that has generated much discussion in the past two years is the idea that rare causal variants arising on the background of common disease-associated haplotypes are actually responsible for producing 'synthetic' GWAS signals.[37] In this scenario, rare causal mutations (not directly tested in GWAS) have occurred (by chance) on haplotypes which share an allele at a common marker that is typed on GWAS (**Figure B5**). Thus, this common allele would show weak association; had the rare variants been tested, their effect sizes might have been larger. It has been

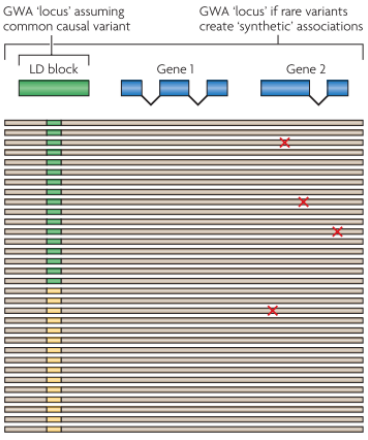


Figure B5: Possible synthetic association at GWAS locus.[32]

especially enticing to consider that the rare variants could be protein-coding (and might thus provide more clear therapeutic hypotheses). At the time this thesis was begun, the discovery and characterization of rare genetic variants had just begun, and the extent to which such synthetic associations might exist at human GWAS loci was not known (this is addressed in Chapters 5-6).

Of course, the common polygenic model originally described by Fisher and the other biometric scientists is also potentially consistent with the findings of GWAS. Although the very existence of 'outlier' GWAS signals may seem incompatible with a model in which variants have infinitesimal effects on phenotype, it has been argued that these signals could still represent the tail of a very large number of common mutations which each increase disease risk by very little. Indeed, it has been demonstrated that weak effects across the full tail of sub-genome-wide significant common variants in aggregate can explain a much larger fraction of disease heritability than just the top signals alone;[36,39,41,45] it is still unclear, however, whether these weak effects represent causal signals or whether they are markers tagging causal variation that might very well have different properties.

Thus, the results of linkage and GWAS studies have been construed in very different ways, to suggest very different properties of genetic architecture. This has occurred, in part, because each study allows investigators only a partial view of the full architecture. A key challenge (especially as genetic studies grow in scope and number) is how to integrate the findings of many studies to draw inference about the underlying 'elephant' that is the genetic architecture (**Figure B6**). A major goal of this thesis was to develop a systematic approach for stating principled hypotheses about genetic architecture and evaluating them against the *entirety* of empirical data.

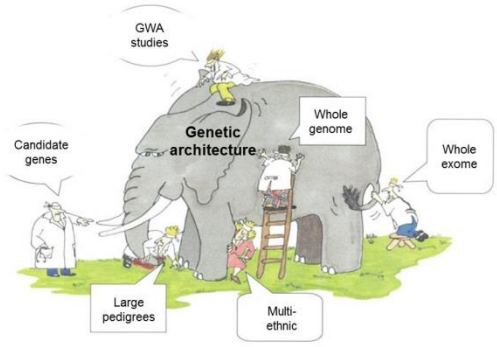


Figure B6: Integrating insights from diverse genetic studies

Next-generation sequencing has recently enabled genetic studies of unprecedented depth

Over the course of this thesis (2009-2013), the cost of sequencing DNA has fallen by more than ten-fold, at a rate that has far surpassed even Moore's Law in the semiconductor industry (**Figure B7**). Next-generation sequencing (NGS) technologies have exemplified the maxim "smaller, faster, and cheaper": gains in cost efficiency have been accompanied by smaller input DNA requirements, lower sequencing time, and improved accuracy. The technical advancements that made this possible are the subject of numerous reviews[55–57], but their impact on human genetics research has been transformative. NGS has made it possible to sequence the genome of many organisms and perform comparative and evolutionary studies; it has enabled high-resolution characterization of the transcriptome (via RNA-seq), and the microbiome; it has spawned a new wave of large-scale biological experiments which leverage sequencing-based read-outs; and of course, it has enabled re-sequencing of the human genome at unprecedented depth in thousands of individuals.

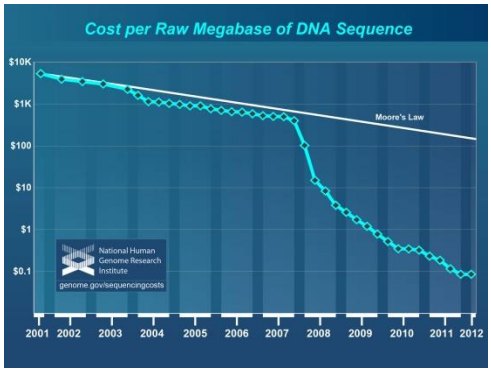


Figure B7: Cost of sequencing 1Mb of DNA, 2001-2012. (Source: NHGRI)

An initial wave of human genetic studies, such as those conducted by the 1000 Genomes Project[54], were not focused on any particular disease trait and instead sought to characterize the full spectrum of human genetic variation (both common and rare) in populations at large. These studies pioneered novel methods for variant detection and genotyping from raw NGS data, and reported on the properties of tens of millions of SNPs, insertions and deletions, and structural variants found to be segregating in human populations. Underscoring the challenge of genotype-phenotype correlation, these studies also documented the finding that the average individual (with no disease phenotype) carries numerous putative loss-of-function variants in annotated genes. Additionally, numerous studies reported on the discovery of large numbers of variants that are rare in the human population; because this has been somewhat inaptly described as an "excess"[50–52] of rare variation

(rather than the naturally expected outcome of human population growth), there has been rampant speculation about the role of these rare sites in human disease phenotypes.

Most recently, a number of next-generation genetic studies have been designed to directly interrogate the genetic basis of human diseases. For type 2 diabetes, a truly incredible set of genetic studies have been performed over the past three years: low pass whole-genome sequencing of 2,800 unrelated European individuals with imputation into >30k individuals, exome sequencing of 13k unrelated individuals of diverse ethnic ancestries and genotyping of low frequency coding variants in up to 80k individuals (via the 'Exome Chip'), and whole-genome sequencing and imputation of ~1,000 individuals from large T2D pedigrees. It is a challenging task to integrate findings from this set of varied studies, and draw cohesive inference about the underlying genetic architecture of T2D. This thesis describes some foundational principles and first steps that will perhaps help guide this effort in coming years.

References

1. Bateson, W. The Methods and Scope of Genetics: An Inaugural Lecture Delivered 23 October 1908. (1908).at <http://www.esp.org>
2. Mendel, G. Experiments in Plant Hybridization. *Read at the February 8th, and March 8th, 1865, meetings of the Brünn Natural History Society* (1865).
3. Garrod, A.E. The Incidence of Alkaptonuria: A Study in Chemical Individuality. *Lancet* **ii**, 1616-1620 (1902).
4. Sturtevant, A.H. *A History of Genetics*. (Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY, 1967).
5. Sturtevant, A.H. The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association. *Journal of Experimental Zoology* **14**, 43-59 (1913).
6. Lander, E.S. & Green, P. Construction of multilocus genetic linkage maps in humans. *PNAS* **84**, 2363-7 (1987).
7. Altshuler, D., Daly, M.J. & Lander, E.S. Genetic mapping in human disease. *Science* **322**, 881-8 (2008).
8. Hanley, J. a "Transmuting" Women into Men: Galton's Family Data on Human Stature. *The American Statistician* **58**, 237-243 (2004).
9. Fisher, R.A. The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Transactions of the Royal Society of Edinburgh* **52**, 399-433 (1918).
10. Falconer, D.S. The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Annals of Human Genetics* **29**, 51-76 (1966).
11. Falconer, D.S. The inheritance of liability to diseases with variable age of onset, with particular reference to diabetes mellitus. *Annals of Human Genetics* **31**, 1-20 (1967).
12. Silventoinen, K. *et al.* Heritability of adult body height: a comparative study of twin cohorts in eight countries. *Twin Research* **6**, 399-408 (2003).

13. Almgren, P. *et al.* Heritability and familiality of type 2 diabetes and related quantitative traits in the Botnia Study. *Diabetologia* **54**, 2811-9 (2011).
14. Reich, D.E. & Lander, E.S. On the allelic spectrum of human disease. *Trends in Genetics* **17**, 502-10 (2001).
15. Chakravarti, A. Population genetics — making sense out of sequence. *Nature Reviews Genetics* **21**, (1999).
16. Hirschhorn, J.N. & Daly, M.J. Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics* **6**, 95-108 (2005).
17. The International HapMap Consortium A haplotype map of the human genome. *Nature* **437**, 1299-320 (2005).
18. The International HapMap Consortium A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851-61 (2007).
19. The Wellcome Trust Case Control Consortium Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661-78 (2007).
20. Sladek, R. *et al.* A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445**, 881-5 (2007).
21. Saxena, R. *et al.* Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science (New York, N.Y.)* **316**, 1331-6 (2007).
22. Scott, L.J. *et al.* A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science (New York, N.Y.)* **316**, 1341-5 (2007).
23. Jostins, L. & Barrett, J.C. Genetic risk prediction in complex disease. *Human Molecular Genetics* **20**, R182-8 (2011).
24. Talmud, P.J. *et al.* Utility of genetic and non-genetic risk factors in prediction of type 2 diabetes: Whitehall II prospective cohort study. *BMJ* **340**, b4838-b4838 (2010).
25. Voight, B.F. *et al.* Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nature Genetics* **42**, 579-89 (2010).
26. Maher, B. The case of the missing heritability. *Nature* **456**, (2008).
27. Botstein, D. & Risch, N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nature Genetics* **33**, 228-37 (2003).
28. Raychaudhuri, S. Mapping rare and common causal alleles for complex human diseases. *Cell* **147**, 57-69 (2011).
29. Lupski, J.R., Belmont, J.W., Boerwinkle, E. & Gibbs, R. a Clan genomics and the complex architecture of human disease. *Cell* **147**, 32-43 (2011).
30. Frazer, K. a, Murray, S.S., Schork, N.J. & Topol, E.J. Human genetic variation and its contribution to complex traits. *Nature Reviews Genetics* **10**, 241-51 (2009).
31. McClellan, J. & King, M.-C. Genetic heterogeneity in human disease. *Cell* **141**, 210-7 (2010).
32. Cirulli, E.T. & Goldstein, D.B. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature Reviews Genetics* **11**, 415-25 (2010).
33. McCarthy, M.I. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics* **9**, 356-69 (2008).
34. Anderson, C. a, Soranzo, N., Zeggini, E. & Barrett, J.C. Synthetic associations are unlikely to account for many common disease genome-wide association signals. *PLoS Biology* **9**, e1000580 (2011).
35. Mitchell, K.J. What is complex about complex disorders? *Genome Biology* **13**, 237 (2012).
36. Purcell, S.M. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748-52 (2009).
37. Dickson, S.P., Wang, K., Krantz, I., Hakonarson, H. & Goldstein, D.B. Rare variants create synthetic genome-wide associations. *PLoS Biology* **8**, e1000294 (2010).
38. Park, J.-H. *et al.* Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nature Genetics* **42**, 570-5 (2010).
39. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics* **42**, 565-9 (2010).
40. Gibson, G. Rare and common variants: twenty arguments. *Nature Reviews Genetics* **13**, 135-45 (2011).
41. Lee, S.H. *et al.* Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nature Genetics* **44**, 247-250 (2012).
42. Wang, K. *et al.* Interpretation of association signals and identification of causal variants from genome-wide association studies. *American Journal of Human Genetics* **86**, 730-42 (2010).

43. Zhu, Q. *et al.* A genome-wide comparison of the functional properties of rare and common genetic variants in humans. *American Journal of Human Genetics* **88**, 458-68 (2011).
44. Wray, N.R., Purcell, S.M. & Visscher, P.M. Synthetic associations created by rare variants do not explain most GWAS results. *PLoS Biology* **9**, e1000579 (2011).
45. Stahl, E. a *et al.* Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. *Nature Genetics* 1-9 (2012).doi:10.1038/ng.2232
46. Goldstein, D.B. The importance of synthetic associations will only be resolved empirically. *PLoS Biology* **9**, e1001008 (2011).
47. Gloyn, A.L. & McCarthy, M.I. Variation across the allele frequency spectrum. *Nature Genetics* **42**, 648-50 (2010).
48. Johansen, C.T. *et al.* Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. *Nature Genetics* **42**, 684-7 (2010).
49. Pritchard, J.K. & Cox, N.J. The allelic architecture of human disease genes: common disease-common variant...or not? *Human Molecular Genetics* **11**, 2417-23 (2002).
50. Coventry, A. *et al.* Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nature Communications* **1**, 131 (2010).
51. Li, Y. *et al.* Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *Nature Genetics* **42**, 969-72 (2010).
52. Keinan, A. & Clark, A.G. Recent Explosive Human Population Growth Has Resulted in an Excess of Rare Genetic Variants. *Science* **336**, 740-743 (2012).
53. Nelson, M. *et al.* An Abundance of Rare Functional Variants in 202 Drug Target Genes Sequenced in 14,002 People. *Science* **337**, 100-104 (2012).
54. The 1000 Genomes Project Consortium A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-73 (2010).
55. Mardis, E.R. Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics* **9**, 387-402 (2008).
56. Metzker, M.L. Sequencing technologies - the next generation. *Nature Reviews Genetics* **11**, 31-46 (2010).
57. Niedringhaus, T.P., Milanova, D., Kerby, M.B., Snyder, M.P. & Barron, A.E. Landscape of next-generation sequencing technologies. *Analytical Chemistry* **83**, 4327-41 (2011).
58. Zeggini, E. *et al.* Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science (New York, N.Y.)* **316**, 1336-41 (2007).

Chapter 1

Forward simulation of population-scale human genetic variation

The genetic architecture of any trait is the result of population genetic processes

Population genetic principles provide a unifying framework in which to consider the genetic basis of complex traits such as T2D. The genetic architecture of any trait has – by necessity – been shaped by the key forces of population genetics: mutation, drift, and selection. Mutations at some (but not all) loci across the genome have the potential to alter disease risk; the size of this 'disease target' will influence the magnitude of effect that individual variants can have on disease phenotype. Genetic drift and gene flow (influenced by human demographic history and migration) cause fluctuations in the frequencies of causal alleles, independent of phenotype effects. Finally, natural selection produces directional changes in the frequencies of alleles that influence evolutionary 'fitness', which is itself a composite of many traits (including, potentially, the disease of interest).

Analytical or simulation-based models have yielded insight into the qualitative dependencies of genetic architecture (usually at a single locus) on subsets of these parameters.[1–7] For example, explosive population growth following a bottleneck can allow even deleterious disease alleles to reach common population frequency.[5] Conversely, strong selection against disease[7], or high mutation rates coupled with mild selection[6], could, in principle, enable rare alleles to explain much of heritability.

To quantitatively investigate the extent to which such models are consistent with emerging data from association studies and population-based sequencing, we performed simulations that enabled granular predictions of genome-wide genetic architecture and study results. Although the number of disease model parameters is potentially without bound, we sought to generate the simplest possible models considering only mutations (of additive effect), genetic drift, and purifying selection. If such simple models produce predictions inconsistent with empirical data, this does not

imply that more complex models could not be consistent. However, if a simple model is consistent, then we can conclude that its features are indeed plausible given current data.

Based on these considerations, we developed a three-stage framework: (1) forward evolutionary simulation to generate multi-locus DNA variation at large scale, (2) mapping of genotype to phenotype under a range of disease models, and (3) *in silico* prediction of genetic study results under each model, followed by comparison to empirical results (**Figure 1.1**). For simplicity, we focused on Northern European populations (in which the majority of human genetic studies have been conducted). Methods and results for each of Steps 1-3 are described in Chapters 1-3.

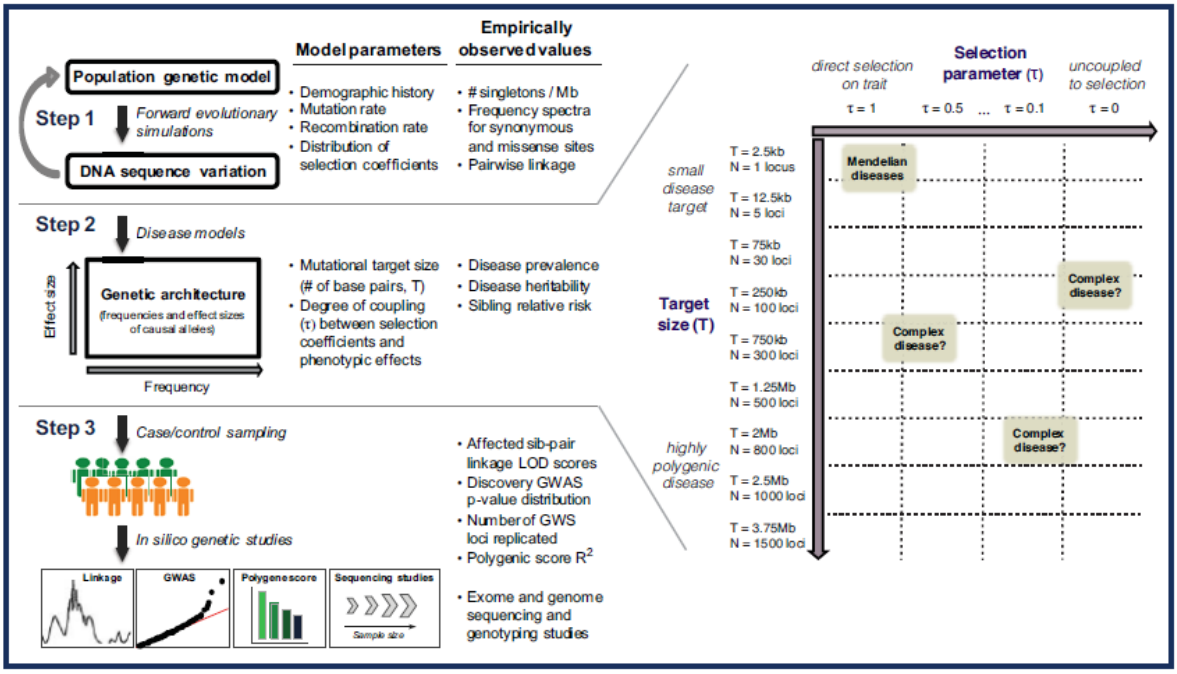


Figure 1.1: Overall framework for simulation and evaluation of disease models for complex traits.

Simulation of DNA sequence variation at population-scale

A pre-requisite for any informative genetic model of human disease is population-scale DNA variation (in hundreds of thousands of samples) – ranging from common polymorphisms to singletons – that is consistent with empirically observed human genetic variation. Since no empirical dataset of this size exists, we opted for a simulation-based strategy.

Simulation tool for forward evolutionary simulations. We chose the publicly available forward evolutionary simulator software ForSim (Lambert et al[8]). Although ForSim is more computationally demanding than coalescent modeling approaches, the ability to simulate rare variation according to a user-specified population demographic history and model of purifying selection was a key advantage. In addition to producing genotype data for all variant sites, ForSim also tracks the evolutionary age, selection coefficient, and haplotype phase information associated with all genetic variants; these additional pieces of information were critical in the downstream implementation of disease models that relied, for example, on assumptions about purifying selection (see Chapter 2).

Evaluation of the simulation tool. Before using ForSim to fit a demographic model, we performed basic diagnostics on simulated data. We ran a simple simulation of N=1,000 individuals (fixed population size) for 10,000 generations at a region of L=50kb. We assumed the mutation rate was u=1.8e-7. We then asked whether the number of novel mutations introduced per generation, the fixation rate of new mutations, and the mean time to fixation were consistent with theoretical predictions under these conditions. Indeed, we found that the mean number of new mutations per generation was ~15, consistent with the expectation $2*N*u*L$ = 18. The expected ratio of the number of sites fixed to the number of sites lost is $(1/(2N)) : (1-(1/(2N))) = 0.0005$. In sample simulated data, we found that ~110 SNPs were fixed, and ~180,000 were lost, yielding a ratio of 0.0006 (consistent with theoretical prediction). Finally, the expected average time to fixation is ~4N, or ~4000 generations. In simulated data, we found that the mean time to fixation (for those sites reaching fixation) was on this order (~5090 generations). Having confirmed that data generated by ForSim met basic diagnostic criteria, we next set out to simulate population-scale DNA variation that was consistent with empirically observed data.

Fitting population genetic parameters to simulate realistic genetic variation. We used ForSim to model the three main processes which determine the spectrum of DNA variation: (a) mutation and recombination, (b) demographic history, and (c) natural selection on segregating alleles. Several models for European population history have been published, but each differs substantially in

assumptions about the mutation rate (μ), ancestral population size (N_a), bottleneck size (N_b), duration (t_e) and rate (r_e) of exponential growth, and modern effective population size (N_e).

We initially tested models of demographic history previously published by Gravel et al (History B; developed based on exome-wide 1000G Pilot data in 63 CEU samples) and Kryukov et al (History C; developed based on a re-sequencing dataset of 58 genes in ~800 individuals). To test the sensitivity of results to bottleneck size, exponential growth curve, and modern effective population size, we tested >20 variants of these histories with different mutation rates (**Table 1.1**).

Table 1.1: Demographic histories evaluated via forward simulation.
'Gravel et al (History B)' and 'Kryukov et al (History C)' represent the original histories published by these groups. 'Hybrid_1 (History A)' represents a demographic history with a bottleneck size of History B, but exponential growth similar to that of History C. History A produced the best fit match to empirical site frequency spectra. Parameters highlighted in blue represent deviations from published histories which were tested directly via forward simulation.

	Demographic history parameters					
	N_A^* (Ancestral population size)	N_B (Bottleneck population size)	t Duration of exponential growth (generations)	r Rate of exp. growth	N_e (Modern effective population size)	u (Mutation rate per bp per generation)
Hybrid_1 (History A)	8,100	2,000	370	1.29%	227,650	2.0 e-8
Hybrid_2	8,100	2,000	920	0.39%	69,852	2.0 e-8
Hybrid_3	7,310 / 14,474	2,000	920	0.39%	69,852	2.0 e-8
Hybrid_4	7,310 / 14,474	2,000	370	1.29%	227,650	2.0 e-8
Gravel_1 (History B)	7,310 / 14,474	1,861 / 1,032	920	0.39%	35,900	2.36 e-8
Gravel_2	7,310 / 14,474	1,861 / 1,032	920	0.39%	35,900	1.2 e-8
Gravel_3	7,310 / 14,474	1,861 / 1,032	920	0.39%	35,900	2.8 e-8
Gravel_4	7,310 / 14,474	1,861 / 1,032	920	0.62%	300,000	2.36 e-8
Gravel_5	7,310 / 14,474	1,861 / 1,032	920	0.67%	500,000	2.36 e-8
Gravel_6	7,310 / 14,474	1,861 / 1,032	1200	0.39%	105,900	2.36 e-8
Gravel_7	7,310 / 14,474	1,861 / 1,032	1500	0.39%	337,475	2.36 e-8
Gravel_8	7,310 / 14,474	6,000 / 4,062	920	0.39%	141,950	2.36 e-8
Gravel_9	7,310 / 14,474	10,000 / 6,712	920	0.39%	234,549	2.36 e-8
Kryukov_1 (History C)	8,100	7,900	370	1.29%	900,000	1.8 e-8
Kryukov_2	8,100	7,900	370	1.29%	900,000	1.2 e-8
Kryukov_3	8,100	7,900	370	1.29%	900,000	2.4 e-8
Kryukov_4	8,100	7,900	370	1.13%	500,000	1.8 e-8
Kryukov_5	8,100	7,900	370	0.88%	300,000	1.8 e-8
Kryukov_6	8,100	7,900	320	1.29%	474,580	1.8 e-8
Kryukov_7	8,100	7,900	250	1.29%	193,730	1.8 e-8
Kryukov_8	8,100	5,000	370	1.41%	900,000	1.8 e-8
Kryukov_9	8,100	2,000	370	1.66%	900,000	1.8 e-8

* Ancestral population sizes separated by '/' indicate the ancestral population size reflect population expansion during the out of Africa event (population size before and after); if not shown, history assumes a constant ancestral population size prior to bottleneck.

To evaluate each demographic history, we used ForSim to forward simulate (with no purifying selection) a 200kb contiguous region in a large (>500K size) population according to the specified history. We repeatedly sampled unrelated individuals (n=63, n=243, and n=1322) from the simulated population for comparison to various empirical datasets of different sample sizes. We then compared the average SFS in simulated samples to the empirical SFS of *synonymous* sites (assumed neutrally evolving) in 1000G exome data (n=63 CEU or n=243 EUR individuals) as well as in exome data generated by the GO-T2D Consortium (n=1322 European samples).

To compare exome-wide data to data at simulated loci, we normalized the mutational target size, using the total length of regions targeted on the exome hybrid capture array (~32.8 Mb total). We assumed that 30% of exonic target would result in synonymous (neutrally evolving) variation if mutated (yielding a synonymous target of ~9.8Mb), while 70% would result in non-synonymous variation (a non-synonymous target of ~22.9Mb).

To account for imperfect sensitivity in variant calling and genotyping in empirical datasets, we estimated that ~95% of all exonic target regions were covered at adequate depth in exome sequencing, and that ~90% of individuals were covered at a given site. We then adjusted simulated site frequency spectra according to these assumptions (see code below) and then compared the corrected simulated SFS under each demographic history to the SFS observed in empirical data.

```
#Code used to correct simulated site frequency spectrum for imperfect empirical sensitivity

#N = number of individuals; also the length of the SFS vector since this extends to MAF=50%
#SFS is site frequency spectrum; number of variants seen at each Minor Allele Count

percent_target_missed = 0.05
percent_individuals_covered = 0.90

N = length(simulated_sfs)
simulated_sfs_corrected <- seq(from=0,to=0,length.out=N)

for (i in 1:N) {
for (j in i:N) {
        p=choose(j,i)*(percent_individuals_covered^i)*(1-percent_individuals_covered)^(j-i)
simulated_sfs_corrected[i] <- simulated_sfs_corrected [i] + p*simulated_sfs[j]
}
}

simulated_sfs_corrected=simulated_sfs_corrected*(1 - percent_target_missed)
```

We confirmed that each demographic history variable (bottleneck size, exponential growth rate, modern effective population size) behaved as expected in forward simulation (**Figures 1.2-1.4**).

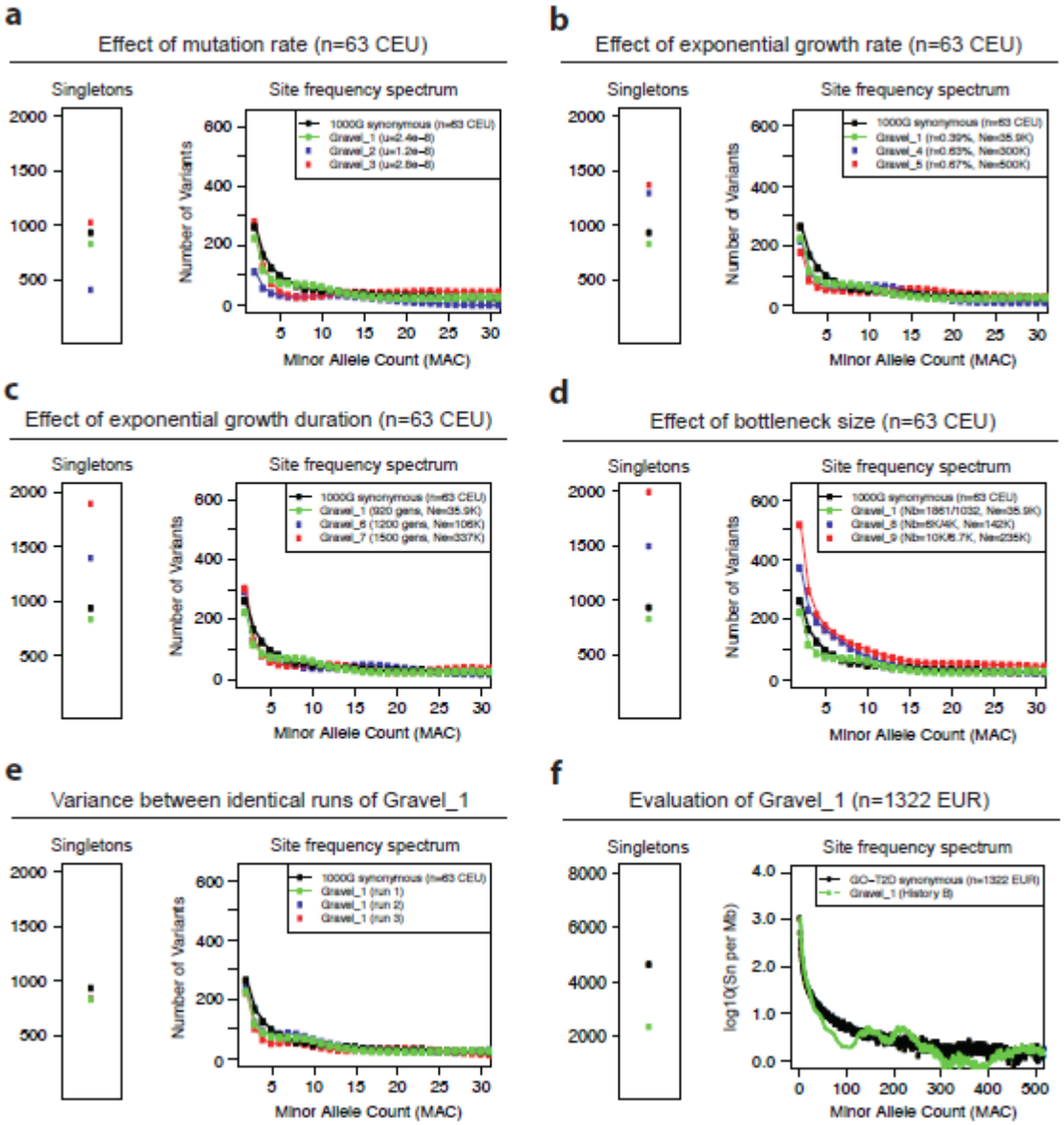


Figure 1.2: Evaluation of Gravel et al demographic history (History B) and variations via forward simulation
The above panels (a-e) evaluate the previously published Gravel et al demographic history and variations of this history (see Table 1 for parameters of each history) against the empirical site frequency spectrum observed empirically in n=63 CEU samples from the 1000G Project. Within each panel, the left plots show number of singletons predicted under each history; right plots show the site frequency spectra from minor allele count of 2 onward. **a)** Effect of varying mutation rate; **b)** Effect of varying rate of exponential growth; **c)** Effect of varying duration of exponential growth; **d)** Effect of varying bottleneck size; **e)** Variance in site frequency spectra produced under identical demographic history (History B); **f)** Comparison of data simulated under History B to the empirical site frequency spectrum observed in a much larger sample size, n=1322 EUR.

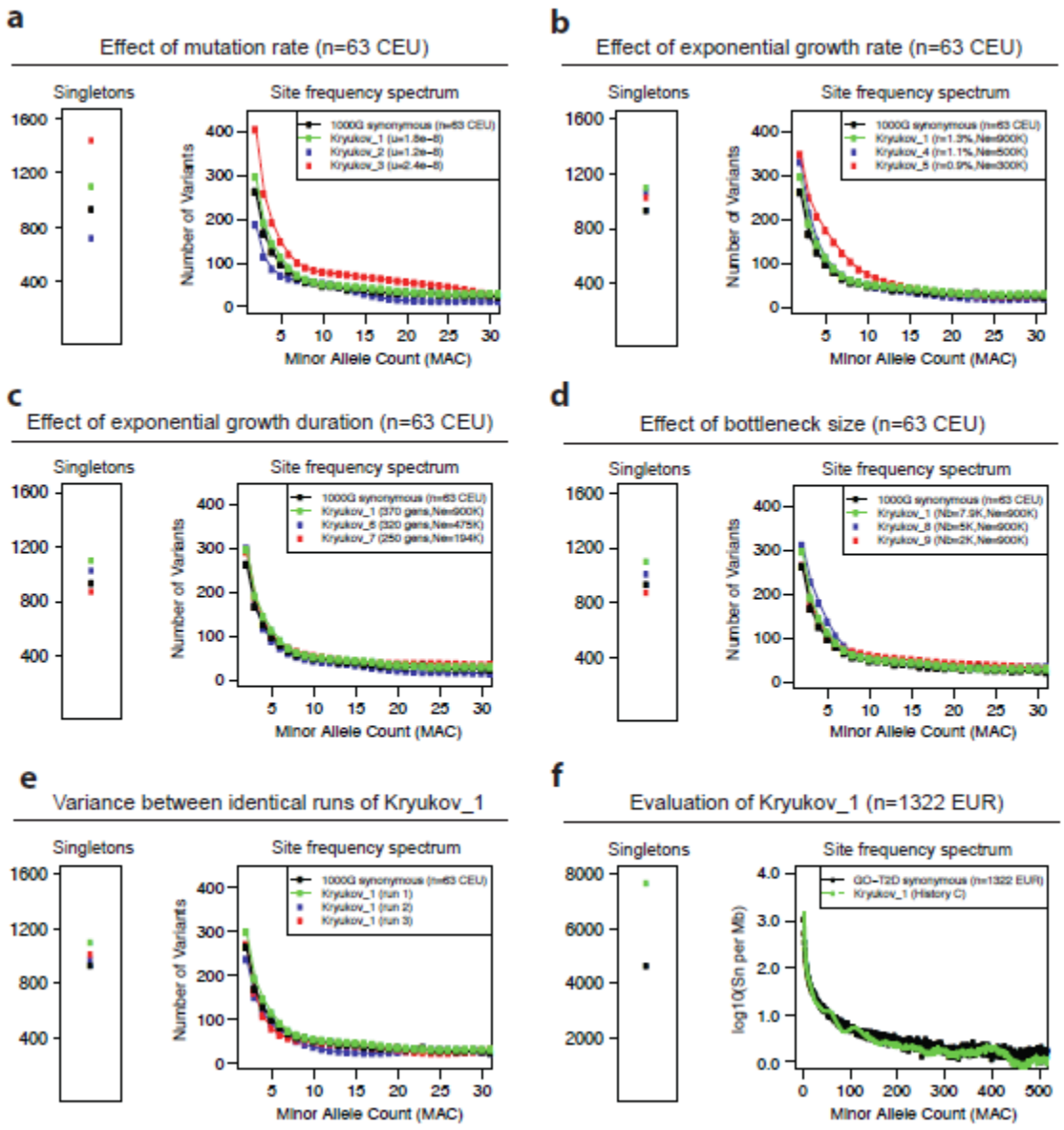


Figure 1.3: Evaluation of Kryukov et al demographic history (History C) and variations via forward simulation
The above panels (a-e) evaluate the previously published Kryukov et al demographic history and variations of this history (see Table 1 for parameters of each history) against the empirical site frequency spectrum observed empirically in n=63 CEU samples from the 1000G Project. Within each panel, the left plots show number of singletons predicted under each history; right plots show the site frequency spectra from minor allele count of 2 onward. **a)** Effect of varying mutation rate; **b)** Effect of varying rate of exponential growth; **c)** Effect of varying duration of exponential growth; **d)** Effect of varying bottleneck size; **e)** Variance in site frequency spectra produced under identical demographic history (History B); **f)** Comparison of data simulated under History B to the empirical site frequency spectrum observed in a much larger sample size, n=1322 EUR.

The above analyses confirmed that the SFS was sensitive to population genetic parameters as expected. A higher mutation rate, as expected, resulted in a greater number of variants at all

frequencies (especially variants of low count). The final effective population size (N_e) heavily impacts the number of segregating rare variants; unsurprisingly, the number of alleles private to a single sampled individual (MAC = 1) is much higher when N_e is large. For comparison, we also evaluated a demographic history published by Schaffner et al (History D; trained only on common variation)[9], as well as a naïve history (History E) of constant population size (N_a=N_b= N_e=10K).

We performed the final evaluation of each history in the large sample size (n=1322; **Figure 1.2f, Figure 1.3f**). We find that History B, due to a small N_e, generates too few rare alleles with MAF < 5%; conversely, History C generates data very close to the empirically observed distribution, but produces an excess of singletons due to a mild bottleneck and large resulting N_e. History D (unsurprisingly, as it was developed from an empirical dataset of only common polymorphisms), is very well-calibrated for common sites but produces insufficient rare sites. Finally, as expected, History E (fixed population size) produces an excess of common sites and far too few rare sites.

We ultimately selected a computationally tractable demographic history which has mixed features of the other histories and which produces a frequency spectrum most similar to that of empirical data (**Figure 1.4**). The demographic history we chose is named "Hybrid_1 (History_A)" in **Table 1.1** and represents a hybrid of previously published histories from Kryukov et al and Gravel et al. This model assumes a mutation rate (μ=2e-8) in line with findings of recent studies.[10–12] It assumes that an ancestral population (N_a=8.1K) underwent a severe bottleneck (to N_b=2K size) approximately 370 generations (t_e) ago, after which time rapid exponential growth (at a rate of r_e=1.3% per generation) occurred to an effective population size N_e=228K (intermediate between that of History B and C). History A closely recapitulates the empirical SFS at both rare and common synonymous sites, in all sample sizes tested (n=63, n=243, and n=1322 individuals; **Figure 1.4**).

We also calibrated a uniform locus-wide recombination rate to match empirically observed pairwise linkage between common variants in data from the 1000 Genomes Project (**Figure 1.5**). A recombination rate of 2 Mb/cM produced the best match to empirical data.

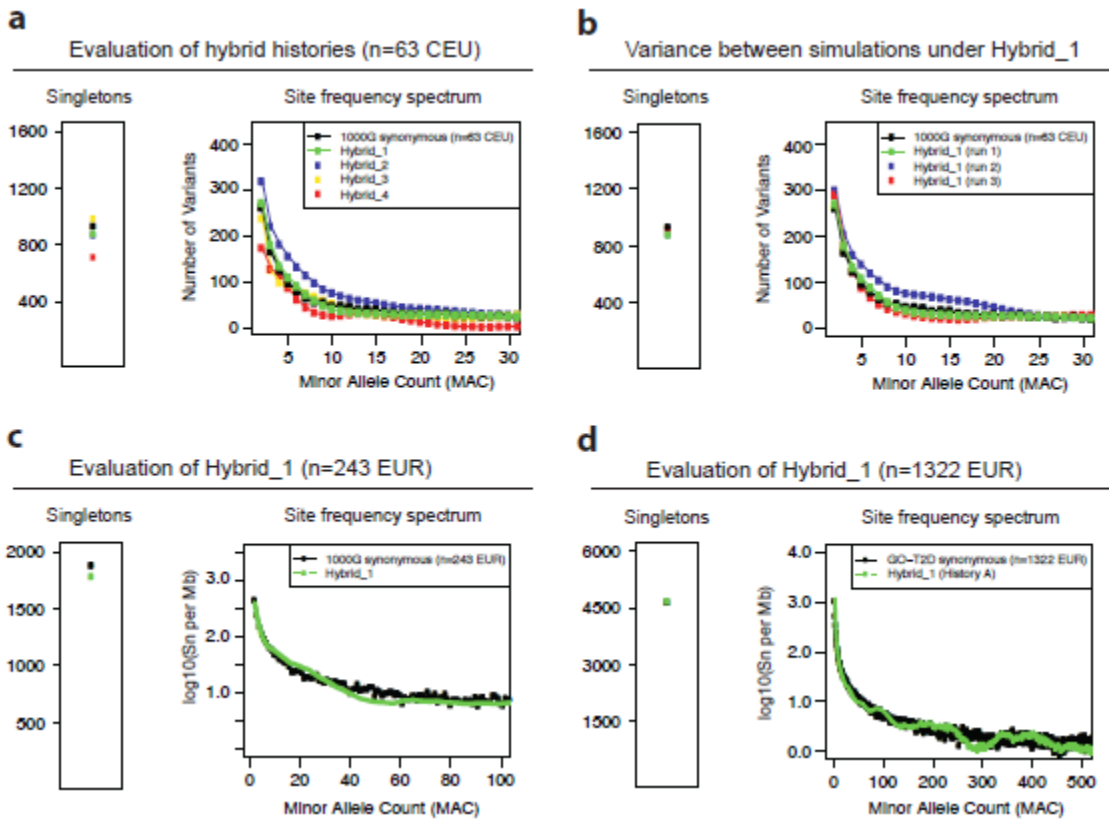


Figure 1.4: Comparison of data simulated under best-fit hybrid history (History A) to empirical datasets.
a) Evaluation of hybrid demographic histories (Table 1) against empirical SFS observed in n=63 CEU samples from 1000G Project. Left plot shows number of singletons predicted under each history; right plot shows SFS from minor allele count of 2 to 30. The history shown in green (Hybrid_1, named "History A") produced the best match to empirical data. **b)** Variance across SFS resulting from independent simulations of an identical demographic history (History A). **c)** Comparison of data simulated under History A to empirical data in a larger sample size, n=243 EUR from the 1000G Project. **d)** Comparison of data simulated under History A to empirical data in a very large sample size, n=1322 EUR from the GO-T2D Consortium**.**

We next fit the distribution of purifying selection on protein-coding mutations by performing forward simulations under History A while applying per-variant selection coefficients drawn from a range of gamma distributions (as in previous reports[13,14], we assumed that ~20% of non-synonymous sites are neutrally evolving). We initially tested the gamma distribution previously published by Kryukov et al. We additionally tested 12 alternate distributions by sampling shape and scale parameters from a logarithmic grid centered at the published values (**Table 1.2**; **Figure 1.6**). Increasing the mean value ($k * \theta$) of the gamma distribution produced, as expected, larger selection coefficients and milder selection, shifting the site frequency spectrum towards more singletons. The best-fit distribution (which produced a SFS closely matching that of *non-synonymous* sites in

empirical data) represents slightly weaker selection than previously published; this may result from our use of a much larger (exome-wide) empirical dataset as compared to the prior study which relied upon genetic variation across only a few (disease-associated) genes.

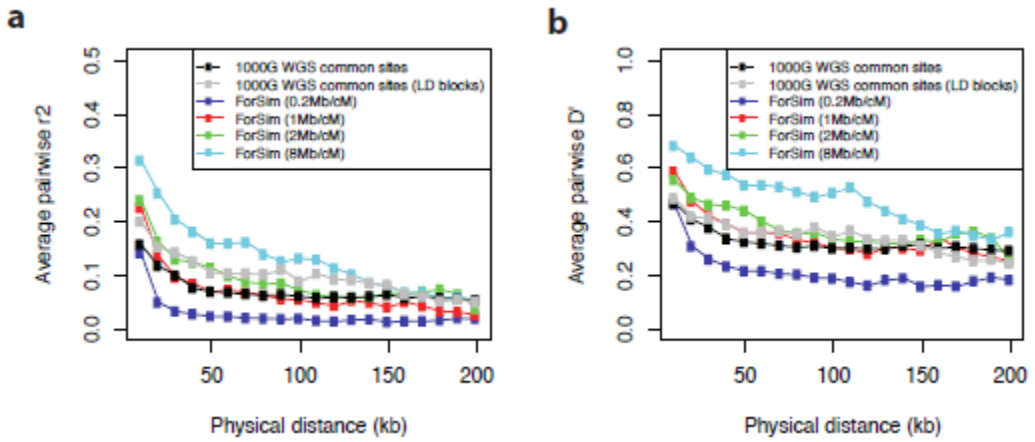


Figure 1.5: Effect of recombination rate on observed linkage disequilibrium between common variants.
a) Average pairwise r2 between pairs of common (MAF>5%) variants in data simulated under History A, with different (uniform) recombination rates. Pairs of common variants are binned by the physical distance between them (x-axis). Empirical data represents common variants on chr1 in 1000G Project WGS data in n=63 CEU individuals. Black line generated from all common variants; grey line generated from only pairs of common variants that lie within the same linkage disequilibrium block (e.g. between two recombination hotspots as measured in the HapMap Project). **b)** Same as **(a)** except shows pairwise D' between common variants.

Table 1.2: Gamma distributions evaluated for distribution of selection coefficients.
Values of theta and k values specifying a gamma distribution of selection coefficients were selected from a logarithmic grid. Pairwise combinations were tested via forward simulation and comparison to the empirical SFS for non-synonymous sites in exome sequencing data. Parameters in the orange row produced the best match to the empirical non-synonymous SFS.

Gamma distribution parameter grid	
θ (scale)	*k (shape)*
0.0010000	0.10000
0.0031623	0.10000
0.0100000	0.10000
0.0316228	0.10000
0.0010000	0.31623
0.0031623	0.31623
0.0100000	0.31623
0.0316228	0.31623
0.0010000	0.56234
0.0031623	0.56234
0.0100000	0.56234
0.0316228	0.56234
0.0010000	1.00000
0.0031623	1.00000
0.0100000	1.00000
0.0316228	1.00000

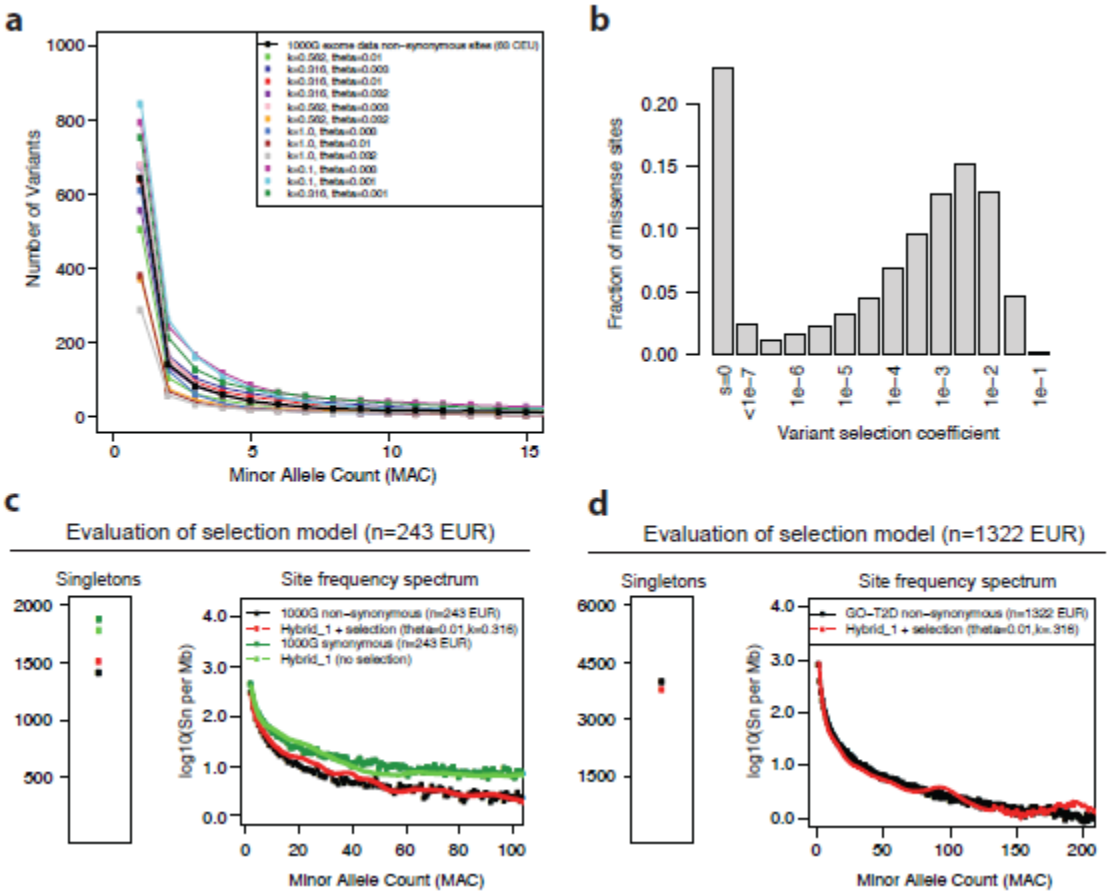


Figure 1.6: Fit of selection coefficient distribution at non-synonymous sites seen in empirical data.
a) Evaluation of different gamma distributions of selection coefficients (see **Table 2**) against empirical site frequency spectrum observed empirically at non-synonymous sites in n=63 CEU samples from the 1000G Project. Based on these data, a gamma distribution with k = 0.316 and theta = 0.01 was selected for forward simulations; shown in **(b)**. **c)** Comparison of non-synonymous data simulated under this selection model (and demographic History A) to empirical non-synonymous data in a larger sample size, n=243 EUR from the 1000G Project. **d)** Comparison of simulated data to empirical data in a very large sample size, n=1322 EUR from the GO-T2D Consortium**.**

We assume all disease loci are under the same distribution of purifying selection (strength comparable to selection at protein-coding changes). This simplifying assumption is likely reasonable for at least a portion of conserved non-coding regulatory elements[15], but future simulations should consider selection distributions matched to different classes of biologically functional loci.

For the present purpose, however, the final set of parameters chosen – mutation rate, recombination rate, demographic history, and selection model – produced simulated data that recapitulated the properties of empirical protein-coding human genetic variation (**Figure 1.7**).

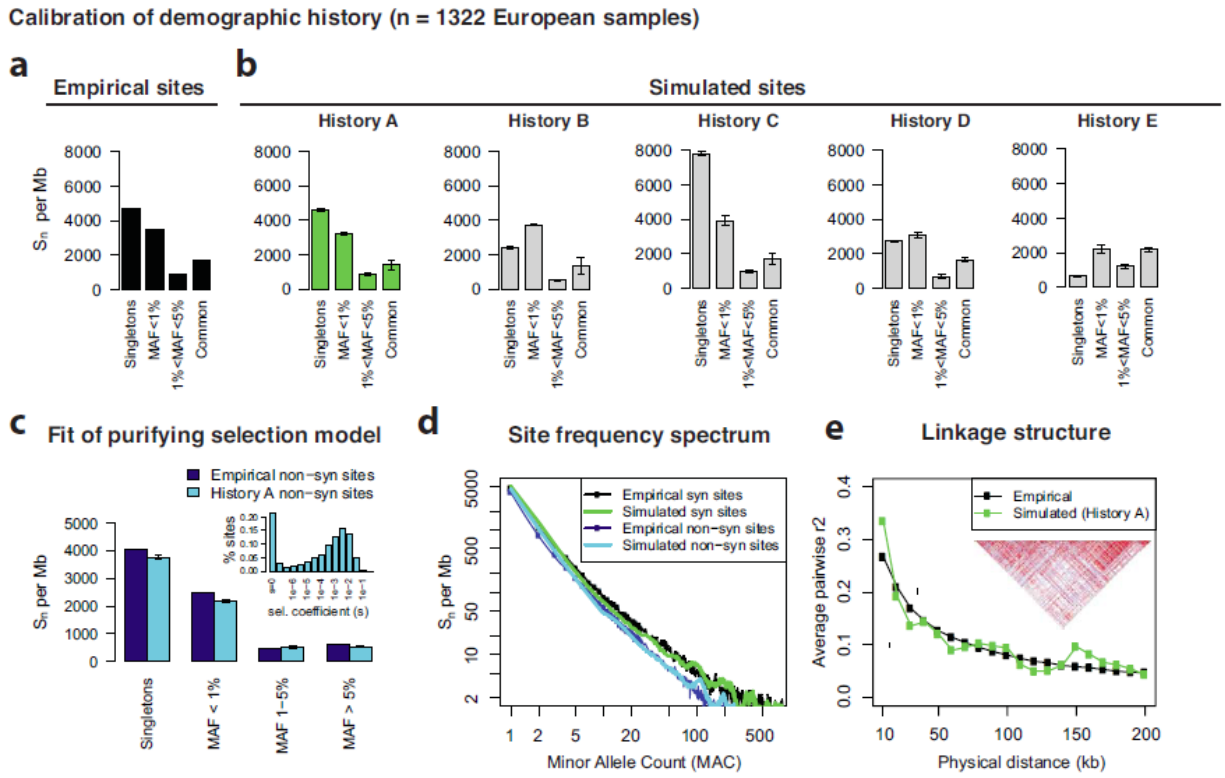


Figure 1.7: Patterns of genetic variation: forward simulated vs. empirically observed
a) Number of singleton, rare (MAF<1%), intermediate frequency (1% < MAF < 5%), and common (MAF > 5%) synonymous sites per Mb of mutational target in empirical data from Go-T2D Consortium, n=1322 European samples. **b)** Number of simulated neutrally evolving sites per Mb under different human demographic histories: A = history chosen in this study (μ=2e-8, N_a=8.1K, N_b=2K, t_e=370 generations, r_e=1.3%, N_e=228K), B = Gravel et al (μ =2.4e-8, N_a=7.3K->14.4K, N_b=1.8K->1.0K, t_e=920 generations, r_e=0.4%, N_e=35.9K), C = Kryukov et al (μ =1.8e-8, N_a=8.1K, N_b=7.9K, t_e=370 generations, r_e=1.3%, N_e=900K), D = Schaffner et al (μ =1.5e-8, N_a=12.5K, N_b=7.7K->540, t_e=350 generations, r_e=0.7%, N_e=100K), E = Fixed 10K population (N_a=N_b= N_e=10K). **c)** Number of non-synonymous (under purifying selection) sites per Mb in empirical data (dark blue) and in forward simulated data (light blue) using chosen demographic history and distribution of selection coefficients (inset). **d)** Full site frequency spectrum (n = 1322 samples) of simulated synonymous (green) and non-synonymous (light blue) sites compared to those in empirical data (black, dark blue). **e)** Average pairwise LD (measured by r^2) as a function of physical distance between frequency-matched common (MAF > 5%) in simulated (green) and empirical (black) data. Linkage structure at a representative 200kb forward simulated locus, as generated in Haploview (inset).

Ratio of missense to silent sites in empirical vs. simulated data under the best-fit parameters.

The ratio of missense to silent sites is a (sample-size dependent) property of empirical sequencing datasets that has been previously reported[16]; this parameter is sometimes used as a quality metric to evaluate variant calling and site annotation in a new sequencing dataset (in fact, at the time the below analyses were performed, they revealed a bug in the annotation software being used by the 1000 Genomes Project). Thus, we explored the properties of this parameter in empirical sequencing datasets as well as in our simulated data (as another test of its similarity to empirical data).

First, we evaluated the distribution of missense vs. silent sites as a function of minor allele count in different sized subsets of the 1000 Genomes Project exome data. Ng et al[16] initially reported on the synonymous and non-synonymous site frequency spectrum observed after exome sequencing at ~50X coverage a sample of 12 individuals from a mix of ethnic backgrounds (4 YRI, 1 CHB, 1 JPT, 6 EUR). To recapitulate this and generate a 'Ng et al-like' dataset, we accessed publicly available (1000G) exome sequence data for 8 of the 12 individuals sequenced by Ng et al; because the other 4 individuals were European-Americans from a private sample, we randomly sampled an additional 4 European samples from the 1000G Project data to obtain a comparable sample of n=12. The numbers of synonymous and non-synonymous sites in this n=12 sample, in n=63 CEU individuals, and in n=822 (multi-ethnic) exomes are shown in **Table 1.3** and **Figure 1.8**.

Table 1.3: Ratio of non-synonymous to synonymous sites in different sized exome sequencing samples
Shown are total (exome-wide) counts of missense (non-synonymous) and silent (synonymous) variants. Cancer genes represent 892 genes that are annotated as related to cancer (http://www.sanger.ac.uk/genetics/CGP/Census).

	Whole exome data					
	All 1000G (N = 822)	1000G CEU only (N = 63)	N = 12 'Ng et al-like'	N = 12 Ng et al 2009	All 1000G (N = 822) *CANCER genes only*	1000G CEU only (N=63) *CANCER genes only*
non-synon total	211,188	36,478	25,961	18,890	5,189	765
synon total	149,453	32,437	28,508	21,201	4,399	876
total # coding SNPs	360,641	68,915	54,469	40,091	9,588	1,641
ratio non-syn : syn sites	1.41	1.12	0.91	0.89	1.18	0.87

In the absence of purifying selection, the missense to silent ratio would be expected to be close to ~2 since there are approximately twice as many exonic positions at which point mutation will produce a missense change as compared to a synonymous change. As seen in **Table 1.3**, however, natural selection removes a number of missense mutations from the population, and makes the observed ratio significantly lower than 2. In larger samples, the ratio is higher because much more rare variation is captured. This class of rare variation is enriched for missense variation because (a) it includes younger mutations that have not yet undergone purifying selection (this class of 'new' mutations are ~2x more likely to be missense than synonymous), and (b) it includes older

missense mutations that have been kept rare due to purifying selection. Interestingly, when we look across only a subset of genes believed to play a role in cancer, the missense to silent ratio is reduced, reflecting an absence of segregating missense sites (presumably due to a higher degree of purifying selection). Thus, this ratio is actually an interesting read-out of selection strength.

Importantly, we confirmed that forward simulated 'exomes' (using the best-fit models of demographic history and purifying selection described earlier in this chapter) recapitulate the observed properties of this missense to silent ratio; simulated synonymous and non-synonymous spectra in a sample of n=63 individuals are shown in **Figure 1.8e**; these data closely resemble the empirical data in **Figure 1.8b**.

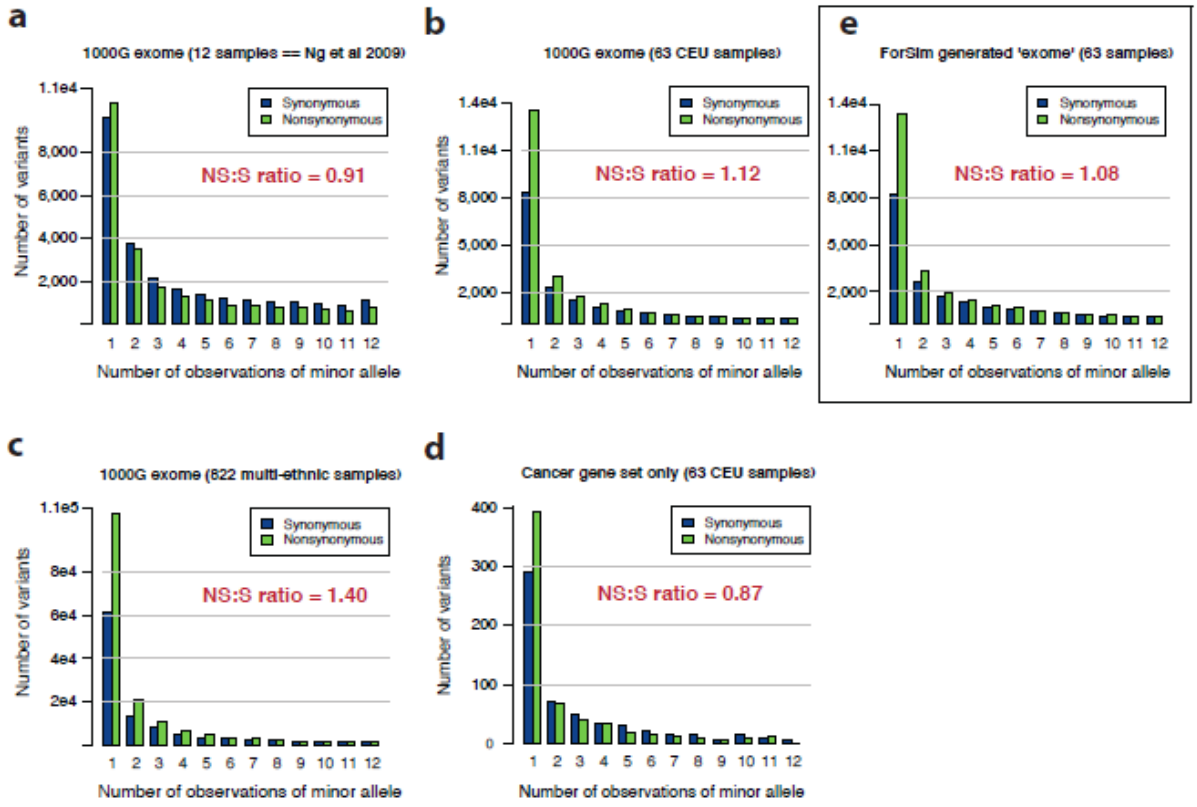


Figure 1.8: Frequency spectra of non-synonymous and synonymous sites in exome sequencing studies of different size, and in forward simulated data.

Simulated disease locus structure. Having confirmed that forward simulated data matches empirical data on a number of dimensions, we next defined the structure of disease loci. ForSim

accepts a flexible, user-defined locus structure. We chose to model disease loci as 'average' protein-coding genes from the RefSeq database. We calculated the median number of exons (8), median total coding length (2.6kb), and median total transcript length (22.2kb) of all unique transcripts in the RefSeq database. Based on this, we model loci in ForSim as a series of 8 exons and 7 introns (alternating), with exons of length 300bp and introns of length 3kb such that the total coding length is 2.4kb and the total transcript length is 23.4kb (**Figure 1.9**). Around each coding locus, we also simulate 100kb of neutral genomic target (in which causal variation does not arise) flanking both sides of the gene to enable downstream genetic association studies with markers in a large window around causal coding variants. It is worth noting that although we have only simulated protein-coding genes in this study, this assumption can be easily modified in future work to model other classes of functional elements (e.g. non-coding regulatory regions) at which a different (perhaps weaker) signature of purifying selection may exist.

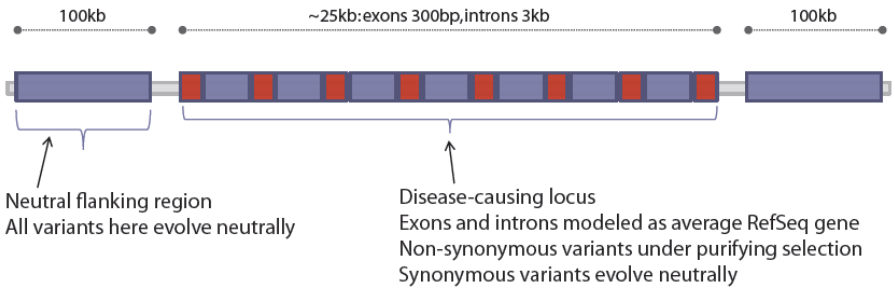


Figure 1.9: Structure of disease loci simulated in ForSim.

Even though the gross structure of all loci generated in ForSim is identical, significant locus heterogeneity is generated by the stochastic nature of the evolutionary process; the number of coding genetic variants segregating in a sub-sampled population of ~10K individuals per locus ranges from 40 to 85, and the number of variants to which we ultimately assign phenotypic effects (see Chapter 2) also varies substantially across individual loci (**Figure 1.10**).

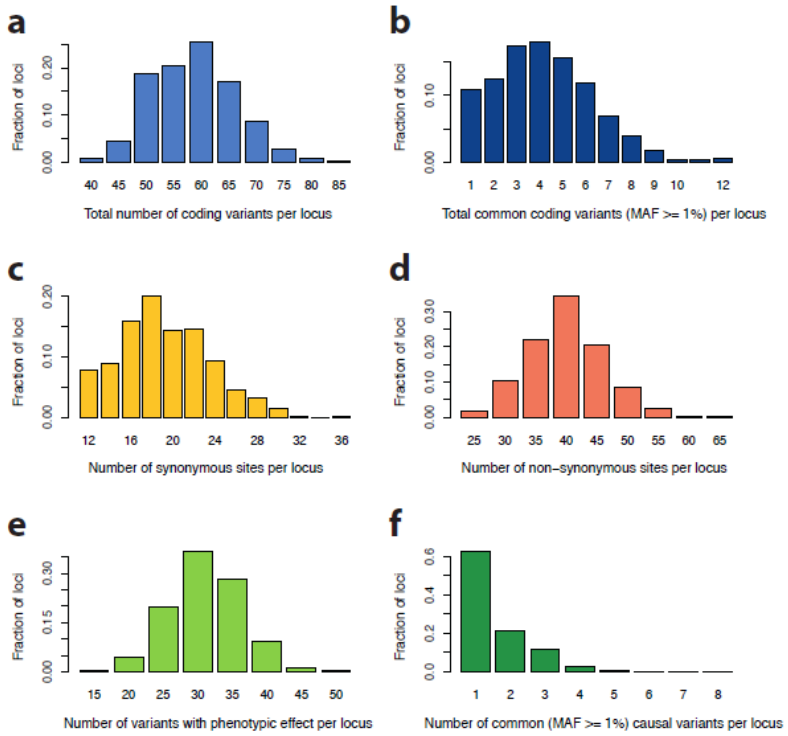


Figure 1.10: Heterogeneity in sequence variation and phenotypic effects across simulated disease loci arising from stochastic forward evolution.

Each panel above is a histogram of simulated loci; in each case, loci are binned by a different feature of sequence variation. Although all loci are simulated with the same physical structure, stochastic forward evolution generates substantial diversity across loci. Simulated loci differ in the number of total coding variants **(a)**, the number of common coding variants **(b)**, the number of synonymous sites **(c)**, the number of non-synonymous sites **(d)**, the number of variants to which phenotypic effect is assigned **(e)**, and the number of common variants to which phenotypic effect is assigned **(f)**.

Finally, as an illustration of the value of empirically calibrated population-scale simulations, we examined the relationship between the frequency of an allele and its deleteriousness (as measured by the selection coefficient). There has been recent speculation, based on observation of abundant rare variation in sequencing studies[17,18], that deleterious alleles are rare, and rare alleles are deleterious; this has in turn suggested support for rare variant models of disease.[19] Simulated sequence data enabled us to test this hypothesis. We thus asked two questions: under the best-fit model of selection, what fraction of deleterious non-synonymous sites is rare? And what fraction of rare non-synonymous sites is deleterious? These quantities can be described as *P(RID)* and *P(DIR)*, respectively, where *R* is the outcome that a non-synonymous variant is rare and *D* is the outcome that the variant has a deleterious effect on fitness (defined here as $s>0.0005$, representing ~50% of all missense sites, as seen in **Figure 1.6b**).

Evaluating *P(RID)*, we find that the vast majority (>90%) of all deleterious non-synonymous variants are indeed rare (MAF< 0.1% in n=10K samples), with the fraction increasing for more deleterious variants (**Figure 1.11**). However, the converse is not true: less than 50% of even extremely rare missense variants (seen 1-2x in 10K samples) are deleterious (**Figure 1.12**), with the remainder essentially neutral (consistent also with recent empirical findings[18]).

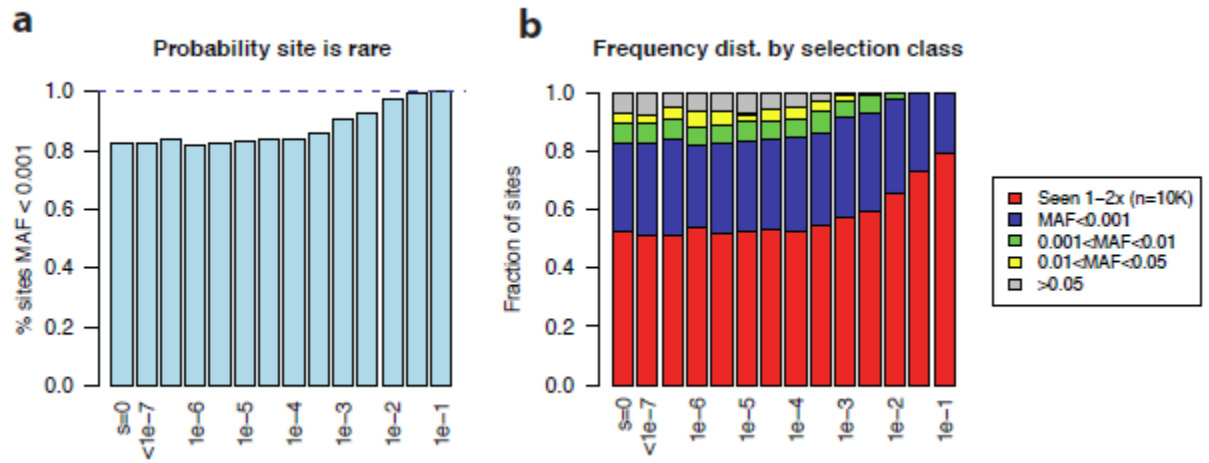


Figure 1.11: Frequency of simulated variants binned by selection pressure.
Under best-fit model of selection, simulated variants are binned by their selection coefficients. **(a)** Fraction of sites in each bin with MAF <=0.1%; **(b)** Distribution of variant frequencies within each bin.

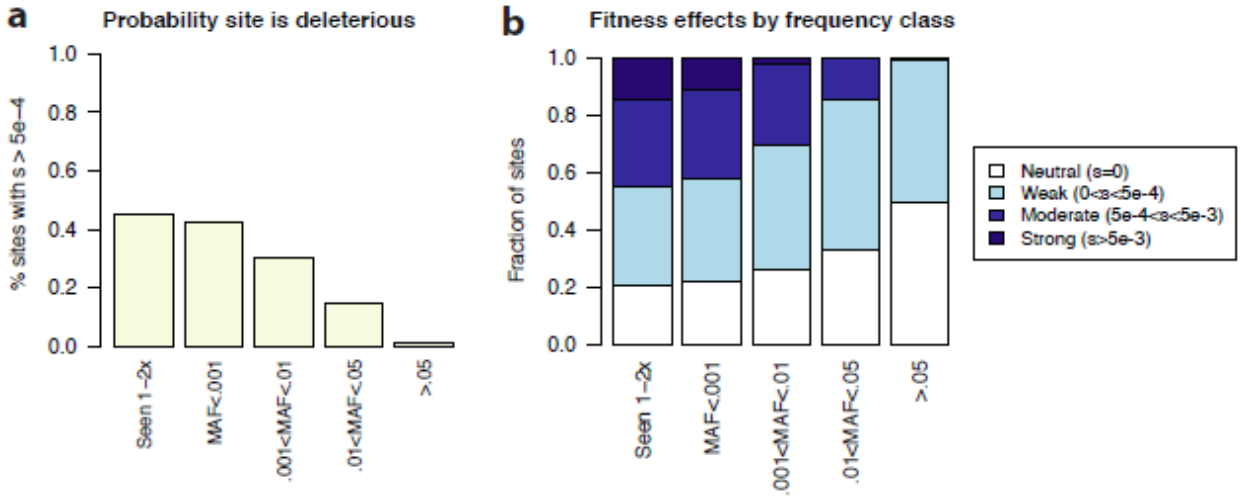


Figure 1.12: Deleteriousness of simulated variants binned by minor allele frequency.
Under best-fit model of selection, simulated variants are binned by their minor allele frequency (x-axis) as measured in a sample of 10K individuals. **(a)** shows fraction of sites in each bin with s >=5e-4 (classified here as 'deleterious'; as seen in Fig 1.11, this represents ~50% of all missense variants, and is thus a lenient definition of deleterious); **(b)** shows the distribution of fitness effects of variants within each minor allele frequency bin. Fewer than half of even very rare sites seen only 1-2X in 10K samples are expected to be deleterious.

This modest probability results from the fact that *P(DIR) = P(RID)* P(D)/P(R)*, and *P(R)* is very high (simply a property of the site frequency spectrum resulting from human population growth). Indeed, the majority of *all* variants are rare due to their recent age rather than any effect on fitness; *P(R)* is >80% even for neutrally evolving non-synonymous sites with no fitness impact (**Figure 1.11**). Thus, most rare variants are simply new (**Figure 1.13**), and these simulations demonstrate that it would be inappropriate to infer functional consequence or historical selection pressure based on a variant's frequency alone.

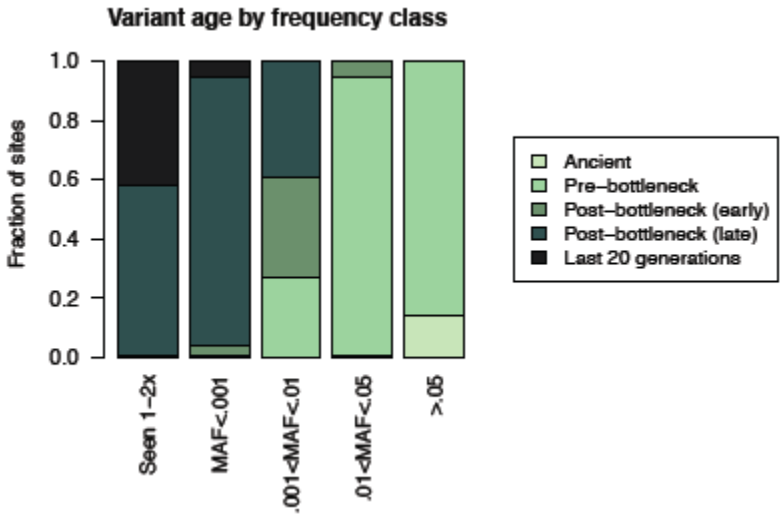


Figure 1.13: Age distribution of simulated variants binned by minor allele frequency.
Under best-fit model of selection, simulated variants are binned by their minor allele frequency (x-axis) as measured in a sample of10K individuals. The distribution of variant age in each bin is shown above.

Collectively, these data demonstrate that variation that has been empirically observed in recent exome studies[20–23] can be robustly simulated in very large sample sizes using a few, simple population genetic parameters (these are available as a ForSim configuration file for use in future studies; see **Appendix A1**). It is important to note that these parameters represent a method by which to produce data that looks realistic and that is calibrated to empirical data; they do not represent inference about the actual demographic history of human populations. Nonetheless, we anticipate that the methods described here may be useful in a variety of settings where simulated data can be informative; to learn properties of sequence variation (as we did in **Figures 1.11-1.13**), to train and optimize novel analytical methods (as we do in Chapter 4), or to evaluate different

genetic study designs. With these simulated genotype data in hand, we next turned to the question of how to map genotype to phenotype and model complex disease genetic architecture.

References

1. Price, A.L. *et al.* Pooled association tests for rare variants in exon-resequencing studies. *American Journal of Human Genetics* **86**, 832-8 (2010).
2. King, C.R., Rathouz, P.J. & Nicolae, D.L. An evolutionary framework for association testing in resequencing studies. *PLoS Genetics* **6**, e1001202 (2010).
3. Browning, S.R. & Thompson, E. a Detecting rare variant associations by identity-by-descent mapping in case-control studies. *Genetics* **190**, 1521-31 (2012).
4. Thornton, K.R., Foran, A.J. & Long, A.D. Properties and Modeling of GWAS when Complex Disease Risk Is Due to Non-Complementing, Deleterious Mutations in Genes of Large Effect. *PLoS Genetics* **9**, e1003258 (2013).
5. Reich, D.E. & Lander, E.S. On the allelic spectrum of human disease. *Trends in Genetics* **17**, 502-10 (2001).
6. Pritchard, J.K. & Cox, N.J. The allelic architecture of human disease genes: common disease-common variant...or not? *Human Molecular Genetics* **11**, 2417-23 (2002).
7. Eyre-Walker, A. Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. *PNAS* **107**, 1752-6 (2010).
8. Lambert, B.W., Terwilliger, J.D. & Weiss, K.M. ForSim: a tool for exploring the genetic architecture of complex traits with controlled truth. *Bioinformatics* **24**, 1821-2 (2008).
9. Schaffner, S.F. *et al.* Calibrating a coalescent simulation of human genome sequence variation. *Genome Research* **15**, 1576-83 (2005).
10. Nachman, M.W. & Crowell, S.L. Estimate of the mutation rate per nucleotide in humans. *Genetics* **156**, 297-304 (2000).
11. Kondrashov, A.S. Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. *Human mutation* **21**, 12-27 (2003).
12. Conrad, D.F. *et al.* Variation in genome-wide mutation rates within and between human families. *Nature Genetics* **43**, 712-714 (2012).
13. Ahituv, N. *et al.* Medical sequencing at the extremes of human body mass. *American Journal of Human Genetics* **80**, 779-91 (2007).
14. Kryukov, G.V., Shpunt, A., Stamatoyannopoulos, J.A. & Sunyaev, S.R. Power of deep, all-exon resequencing for discovery of human trait genes. *PNAS* **106**, 3871-6 (2009).
15. Ward, L.D. & Kellis, M. Evidence of Abundant Purifying Selection in Humans for Recently Acquired Regulatory Functions. *Science* **1675**, (2012).
16. Ng, S.B. *et al.* Targeted Capture and Massively Parallel Sequencing of Twelve Human Exomes. *Nature* **461**, 272-276 (2009).
17. Cirulli, E.T. & Goldstein, D.B. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature Reviews Genetics* **11**, 415-25 (2010).
18. Nelson, M. *et al.* An Abundance of Rare Functional Variants in 202 Drug Target Genes Sequenced in 14,002 People. *Science* **337**, 100-104 (2012).
19. McClellan, J. & King, M.-C. Genetic heterogeneity in human disease. *Cell* **141**, 210-7 (2010).
20. Coventry, A. *et al.* Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nature Communications* **1**, 131 (2010).
21. Johansen, C.T. *et al.* Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. *Nature Genetics* **42**, 684-7 (2010).
22. Li, Y. *et al.* Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *Nature Genetics* **42**, 969-72 (2010).
23. Keinan, A. & Clark, A.G. Recent Explosive Human Population Growth Has Resulted in an Excess of Rare Genetic Variants. *Science* **336**, 740-743 (2012).

Chapter 2

Modeling complex disease genetic architecture in populations

Simple models of complex disease

Having simulated extensive genotype data at loci that resemble human protein-coding genes in a large population, the next key challenge was to develop a principled approach for assigning phenotype in this population. We define *disease model* as the mapping between simulated genotypes and individual phenotypes; it describes the evolutionary basis of a genetic disease.

Under an additive liability threshold model, the relationship between genotype and phenotype is controlled by (a) the number of disease variants carried by an individual; (b) the effects on disease of each causal variant (these effects may or may not be related to the variant's selection coefficient); (c) the magnitude of non-genetic (e.g., environmental) influences; and (d) the liability threshold above which disease ensues. By modulating these levers, it is possible to model a principled distribution of causal variant frequencies and effect sizes rather than specify them *ad hoc*.

For a complex disease, the prevalence (~8% for T2D[1]) determines the liability threshold, and the heritability (~45% for T2D, estimated from family studies[2]) determines the relative magnitude of genetic (as compared to environmental) effects (**Figure 2.1**). To model the number and effect sizes of causal variants, we specify only two variable parameters: the mutational target size (T) and the selection parameter (τ), a measure of coupling between purifying selection and phenotypic effects.

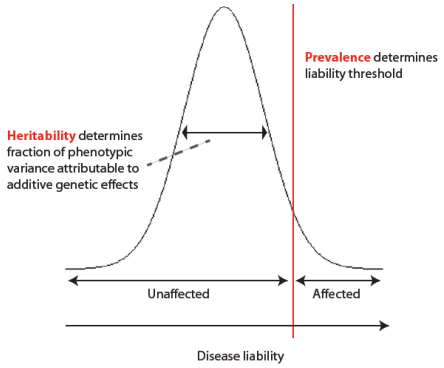


Figure 2.1: Liability threshold model of complex disease

The number of disease variants carried by an individual is determined by the disease mutational target size (*T*), or the sum total of nucleotides that, if mutated, would influence risk of disease. In this work, we assume that causal mutations occur only at sites under evolutionary constraint similar to that at non-synonymous changes under some purifying selection; thus only protein-coding loci and conserved non-coding regions[3,4] (collectively spanning ~10%, or ~300Mb, of the human genome[5,6]) contribute to disease risk. We simulated models with *T* ranging from 75kb to 3.75Mb, corresponding to 0.02%-1.2% of this constrained genome sequence. To model linkage between variants at structurally contiguous genomic regions, we grouped the disease target into 'loci' (*N*=30, 100, 300, 500, 800, or 1500 causal loci in each model). Each locus contains 2.4kb of functional target (under selection) flanked by neutrally evolving regions (as in **Figure 1.9**).

The selection parameter (τ) describes the coupling between the strength of purifying selection and the phenotypic effect for each causal mutation. For lethal Mendelian diseases manifesting prior to reproduction, the coupling of phenotypic effect to purifying selection is clear (disease-causing variants are under clear, direct purifying selection which keeps them rare in the population). But in the general case, where disease manifests after reproduction, or where there is pleiotropy, the relationship between selection and disease is less clear.[7] In most cases, overall fitness results from multiple phenotypes (not a single trait), and each causal mutation may have multiple effects. Thus, there exist a range of possible mappings between a variant's selection coefficient (s) and its effect on a given disease (g). To represent this mapping using a single parameter (τ), we model the additive (disease-specific) effect of each variant as: $g \sim s^{\tau}$ (see Eyre-Walker[8]). We performed simulations with τ = 0, 0.1, 0.3, 0.5, and 1. Where τ = 1 ('tightly coupled'), variants with large effects on fitness have large effects on disease (e.g., there is direct selection against the disease phenotype). Where τ = 0 ('uncoupled'), there is no relationship between selection coefficients of causal mutations and their impact on the disease of interest (**Figure 2.2**).

These two parameters (T and τ) are used to map genotype to phenotype for each individual in the simulated population using the following procedure. First, each individual variant is assigned

an additive phenotypic effect based on the selection coefficient under which it evolved and the selection parameter assumed in the disease model:

$$g = s^{\tau} * (1 + \epsilon), \qquad \text{(Equation 1)}$$

where g is the variant's additive effect on phenotype, s is the selection coefficient, and ϵ is a random variable drawn from a normal distribution.

We assume that all genetic effects are additive, and further assume that there are no interactions between individual variants (no epistasis). We thus assign each individual a total 'genetic phenotype' G by simply summing the effects across all variants for which an individual carries the novel (non-ancestral) allele, across all disease-causing loci:

$$G_k = \sum_{j=1}^{N} \sum_{i=1}^{m} g_{ij}, \qquad \text{(Equation 2)}$$

where g_{ij} is the effect of the ith variant at the jth gene locus at which individual k carries a disease-causing allele. The target size is represented by N, the total number of causal loci over which the genetic effects are summed. Given a population distribution of genetic phenotypes G, we transform these to *z*-scores such that $Var(G) = 1$ by applying the transformation:

$$G_k^z = (G_k - mean(G_k))/stdev(G_k) \qquad \text{(Equation 3)}$$

We then assign environmental phenotypes E, are each drawn from a normal distribution $N(0,1)$ and are weighted by a constant factor b, to each individual to obtain a total phenotype P:

$$P_k = G_k^z + bE_k, \qquad \text{(Equation 4.1)}$$

Assuming that genetic effects are independent of environmental effects, the co-variance of these population distributions is zero and so the total phenotypic variance is:

$$Var(P) = Var(G^z + bE) = Var\,(G) + b^2 Var(E) + Cov(G, E)$$

$$Var(P) = Var(G^z) + b^2 Var(E) \quad \text{(Equation 4.2)}$$

The weighting constant b is constrained by the empirically observed heritability of the disease, h such that:

$$h = \frac{Var(G^z)}{Var(P)} = \frac{Var(G^z)}{Var(G^z)+b^2 Var(E)} \quad \text{(Equation 4.3)}$$

Since $Var(G^z) = 1$ and $Var(E) = 1$, we find b as a function of the parameter h:

$$b = \sqrt{\frac{(1-h)}{h}} \quad \text{(Equation 4.4)}$$

Thus, each individual's total additive phenotype is calculated as:

$$P_k = G_k^z + \sqrt{(1-h)/h} * E_k \quad \text{(Equation 4.5)}$$

Finally, we apply a liability threshold model for categorical disease phenotypes. Given the final population distribution of P, we use the disease prevalence p to calculate a threshold t above which an individual with phenotype P_k is affected by disease. Case and control status in the population, A, is assigned such that the total fraction of the population that is affected with disease is p.

$$Categorical\ disease\ status = A_k = \begin{cases} 1\ (control),\ P_k < t \\ 2\quad (case),\ P_k \geq t \end{cases} \quad \text{(Equation 5)}$$

In this way, each individual's genotype is mapped to a binary complex disease phenotype using only two disease model parameters: the target size T and selection parameter τ.

The space of disease models tested

We performed simulations for a range of target sizes and selection parameter values which defined a two-dimensional space of disease models (**Figure 1.1**). Under each disease model, we simulated five fully independent replicates. For each replicate, we randomly sampled N loci from

1500 unique loci simulated in ForSim (as described in Chapter 1), and performed all downstream steps (assigned each variant phenotypic effects, added environmental phenotypes, assigned disease status to individuals) independently.

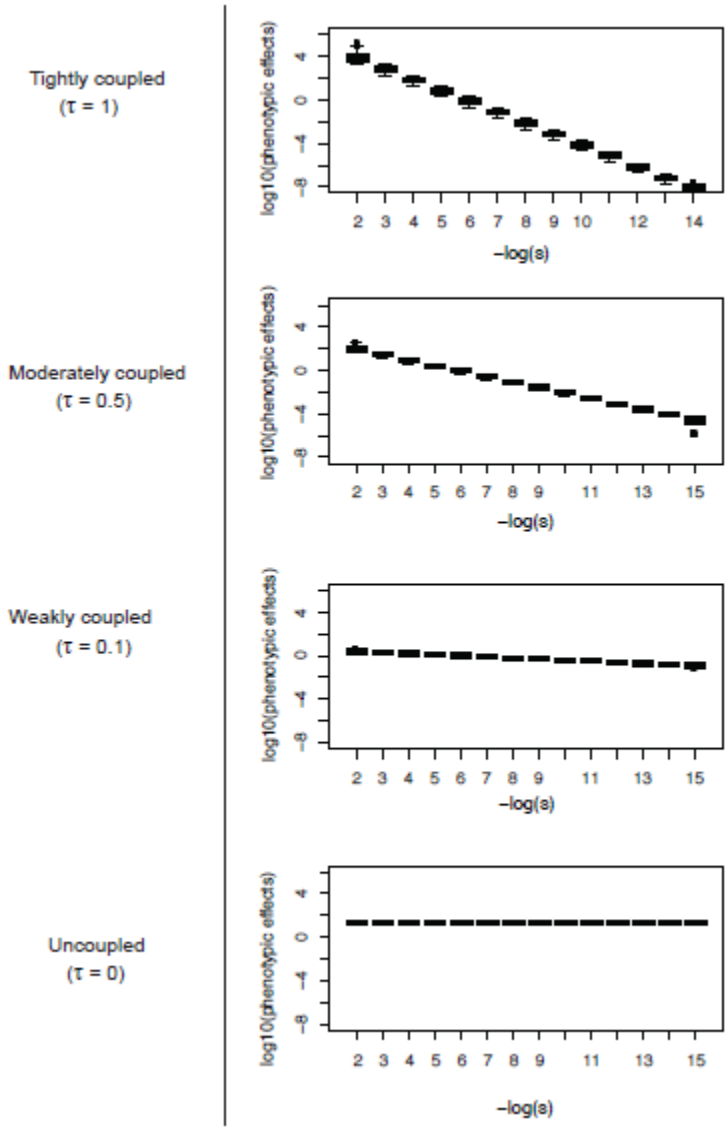


Figure 2.2: Relationship between causal variant selection coefficient and phenotypic effect under different disease models (with varying selection parameter τ).

From each simulated population (in which both genotype and phenotype is now known for each individual), we sampled full-sibling pairs to confirm that the T2D heritability specified under the disease model (h=45%) could be recovered via phenotypic regression and analysis of variance[9]

(**Figure 2.3; Table 2.1**). These analyses were performed using the underlying quantitative trait rather than the dichotomous trait, since heritability was set under a liability threshold model.

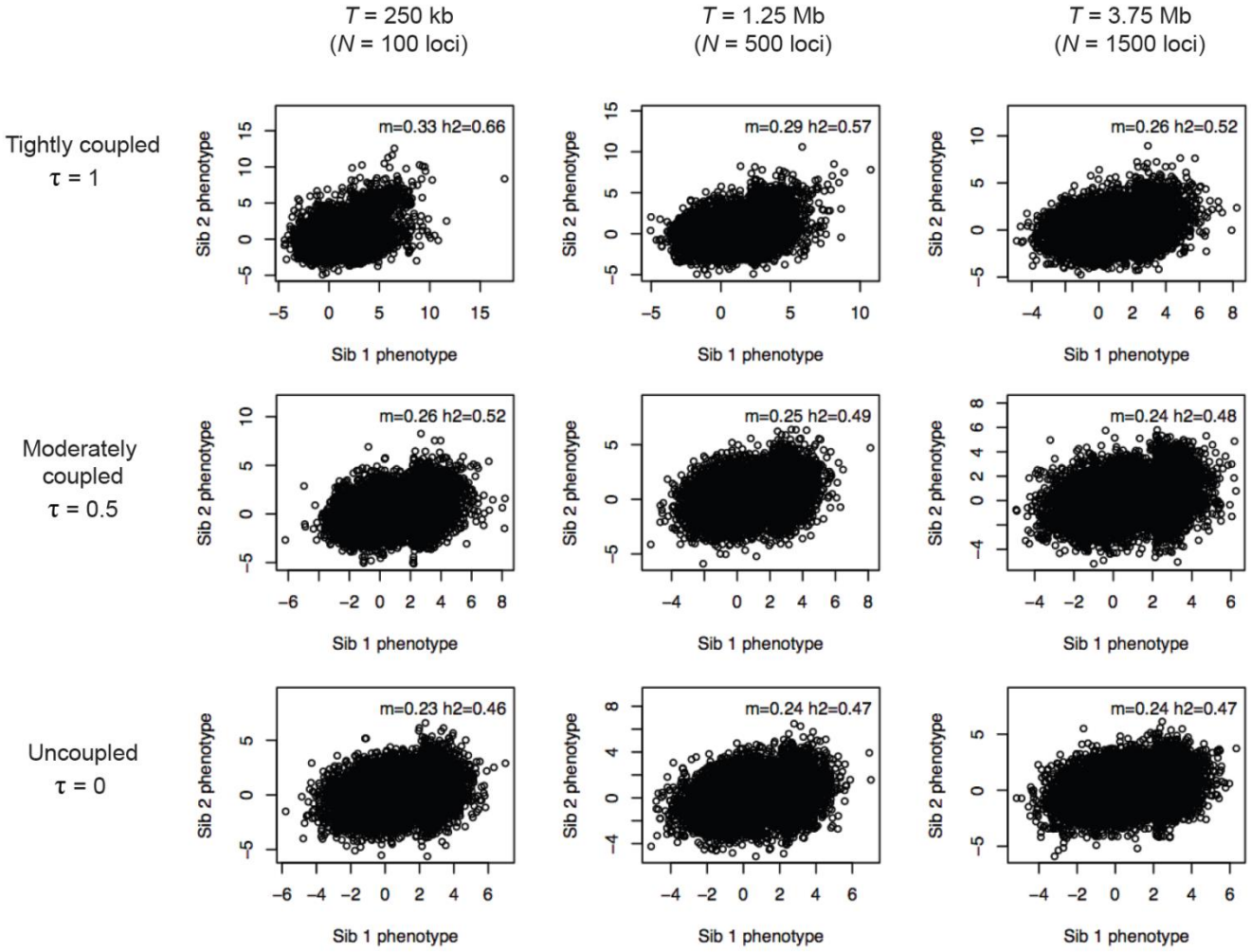


Figure 2.3: Estimation of disease heritability via regression between sibling pairs sampled from populations.

Table 2.1: Estimation of disease heritability via ANOVA performed in sampled full-sibling pairs.

Selection parameter (τ)	Target size (loci)	Mean ANOVA h^2 estimate (across simulation replicates)
τ = 1	250kb (100 loci)	0.45
τ = 1	1.25Mb (500 loci)	0.45
τ = 1	3.75Mb (1500 loci)	0.43
τ = 0.5	250kb (100 loci)	0.46
τ = 0.5	1.25Mb (500 loci)	0.42
τ = 0.5	3.75Mb (1500 loci)	0.45
τ = 0	250kb (100 loci)	0.46
τ = 0	1.25Mb (500 loci)	0.43
τ = 0	3.75Mb (1500 loci)	0.45

In all cases, the standard error = ~0.027 in 2500 families, each with 2 full siblings.

We find that under tightly coupled models ($\tau = 1$), the heritability is slightly over-estimated using linear regression due to concordant outliers carrying mutations of very large effect, but ANOVA confirms that the genetic contribution to trait variance is as expected under all disease models (close to 45%).

Genetic architecture resulting under each disease model

Importantly, the genetic architecture – that is the spectrum of causal variant frequencies and effect sizes – are *outputs* of these disease models and are *not* specified as inputs. The only two parameters we specify are the target size and the selection parameter (T and τ); we hypothesized that varying these would produce genetic architectures with different properties. While we had some analytical expectation for how architecture would vary with each parameter, we next sought to verify this.

We first asked: how do rare and common variant effect sizes compare under different models? We find that (as expected) under tightly coupled (τ=1) models, rare variants (those under strong purifying selection) have much larger effects than common variants, while under uncoupled (τ=0) models, rare and common alleles have comparable additive phenotypic effects (though there is still greater variance around the odds ratios measured for rare variants; **Figure 2.4**). In contrast, the target size does not impact the relative effect sizes of rare and common variants; rather, increases in target size reduce causal variant effects across the entire frequency spectrum. This occurs because T2D prevalence and heritability are fixed, so a larger number of causal variants must be counteracted by smaller per-variant effects (**Figure 2.5**, **Table 2.2**). Notably, under all models, the high prevalence and modest heritability of T2D constrain common (MAF>5%) variants to odds ratios <2, even at relatively small target sizes (e.g. T=75kb).

Next, we asked: how is disease heritability partitioned by allele frequency across the models? The contribution of each causal variant to heritability (population genetic variance) is: $V_a = 2 * (g^2) * (1 - f) * f^{10}$, where g is the variant's additive effect and f is its frequency. Under

tightly coupled (τ=1) models, where g is very large for some rare alleles (often private to cases), the rare class (MAF<1%) collectively explains >90% of heritability. Conversely, under uncoupled models (τ=0), common (MAF>5%) alleles with modest effects (OR<1.2) explain ~95% of heritability (**Figure 2.6**). These relationships hold regardless of the target size.

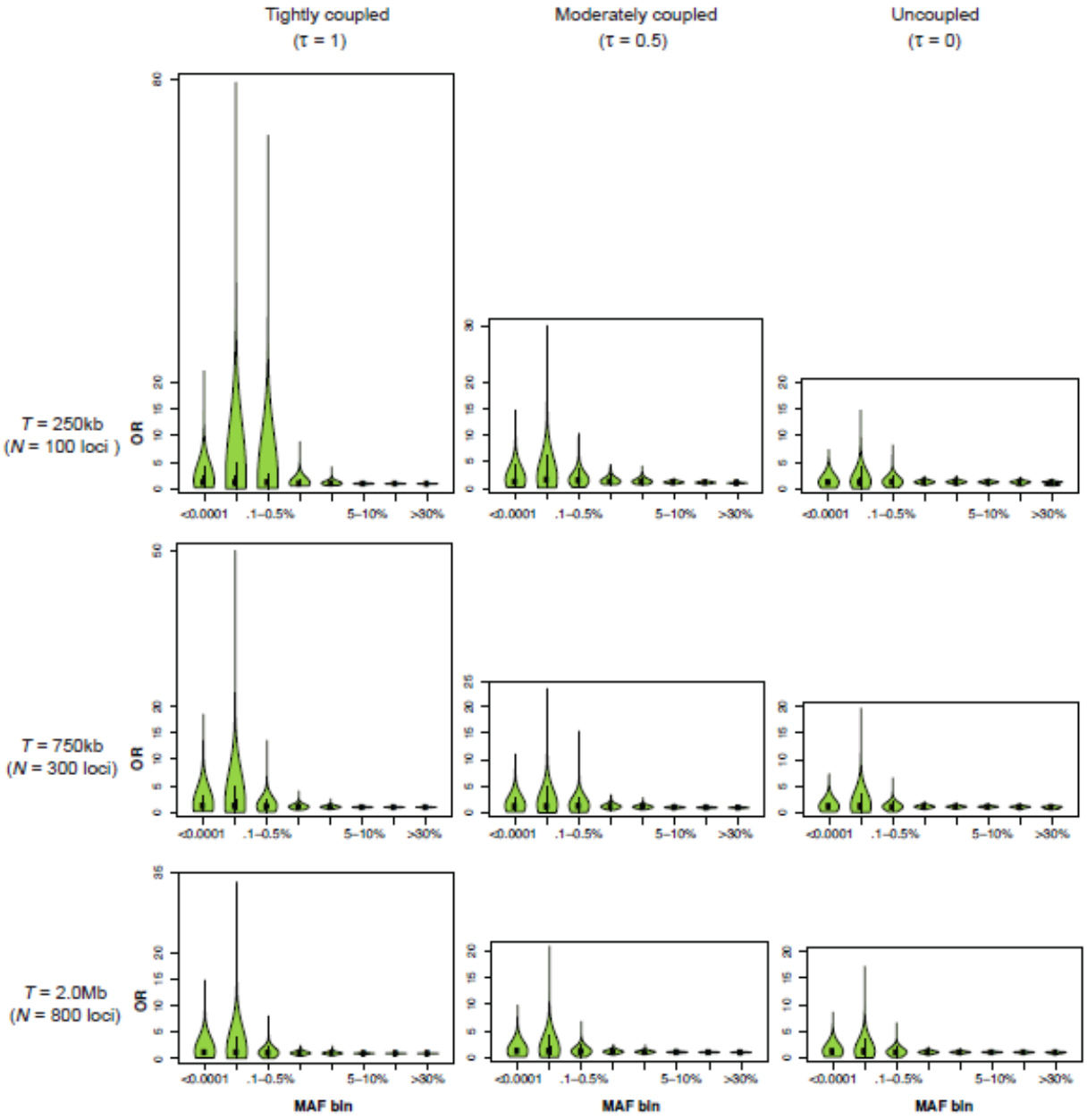


Figure 2.4: Distribution of rare and common causal variant odds ratios (as measured in 10K individuals sampled from simulated populations) under different disease models.

Table 2.2: Total number of segregating causal variants under disease models with varying target size (*T*); shown for tau = 0.5.

Number of segregating causal variants in population				
T (total target)	*N* (number of loci)	# common variants (MAF > 2%)	# intermediate freq. variants	# rare variants (MAF < 0.5%)
75kb	30	29	15	1,946
250kb	100	101	56	6,448
750kb	300	296	164	19,386
1.25Mb	500	498	282	32,334
2.0Mb	800	803	450	51,629
3.75Mb	1500	1516	850	97,008

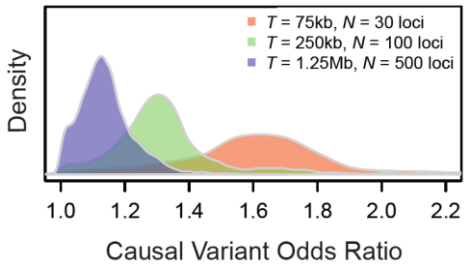


Figure 2.5: Distribution of common (MAF>5%) causal variant odds ratios as a function of target size.

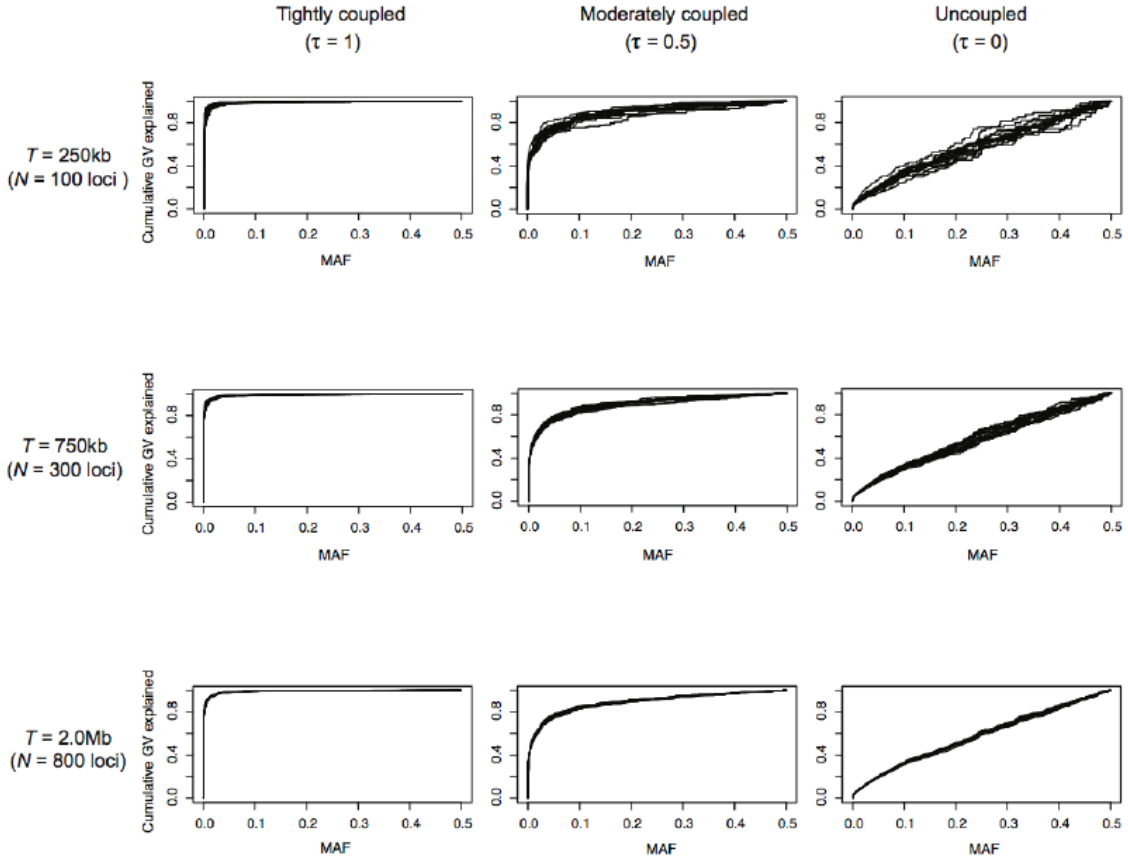


Figure 2.6: Partitioning of population genetic variance by minor allele frequency under different disease models. Each line represents an independent simulation replicate.

Finally, we examined the distribution of variant effects *within each individual* (rather than population-wide) to evaluate the potential of individualized risk prediction. Under each simulated disease model, we asked how many unique risk variants each individual carried (**Table 2.3**). As expected under an additive model, cases carry more disease-causing variants than controls.

Table 2.3: Average number of causal risk alleles *per individual* under disease models with varying target size (*T*); shown for tau = 0.5.

T (total target)	*N* (number of loci)	# common variants (MAF > 2%)		# intermediate freq. variants		# rare variants (MAF < 0.5%)	
		Case	Control	Case	Control	Case	Control
75kb	30	9	8	0.2	0	1	0
250kb	100	30	28	1.4	1	2	1
750kb	300	90	87	4	3	5	4
1.25Mb	500	150	146	7	6	8	6
2.0Mb	800	242	236	11	9	13	10
3.75Mb	1500	462	455	19	18	23	19

While the total number of causal alleles per person depends largely on the target size alone, the *distribution* of additive effects across these causal variants is a function of the selection parameter. From each population, we randomly sampled eight individuals affected with type 2 diabetes, and studied the distribution of additive effects carried by each individual. We find that under tightly coupled models (even if the target size is relatively large), patients with T2D have only few (1-5) high-effect risk alleles (**Figure 2.7**), and these alleles are rarely seen among unaffected individuals.

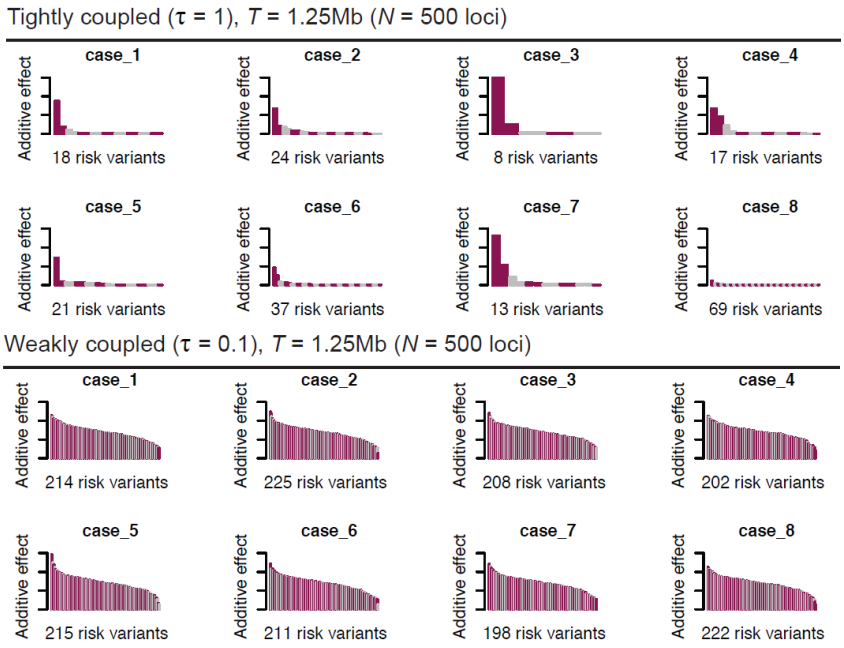


Figure 2.7: Distribution of additive effect sizes per individual patient under different disease models.

Under each model, 8 patients are randomly sampled from the population. Y-axis shows additive effect size of each causal allele in the patient; variants are ordered by descending effect size. For visualization, variants are alternately colored purple or grey in groups of two.

Conversely, under weakly coupled models with similar target size, each patient has hundreds of risk alleles with similar individual and cumulative effect (**Figure 2.7**); moreover, most of these alleles are also commonly observed among controls. Thus, for a given target size, genetic risk prediction will be far more informative (diagnostic for some patients, assuming that effects at rare alleles can be discovered and accurately quantified) if there is strong coupling to selection. This confirms the widely-discussed intuition that, under rare variant models of common disease, data from sequencing studies may greatly enhance clinical risk prediction.

These properties of genetic architecture are perhaps best summarized by visualizing the distribution of *heritability* as a joint function of causal allele frequency and effect size (**Figure 2.8**).

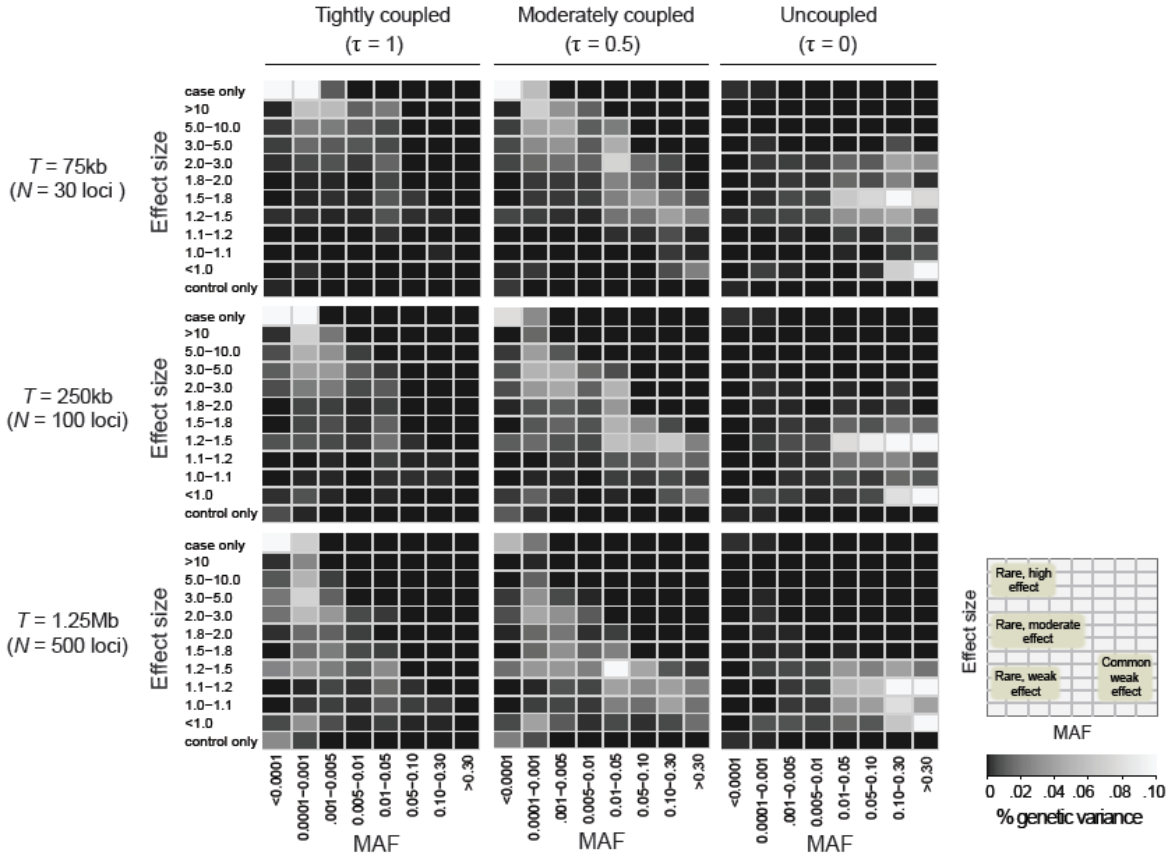


Figure 2.8: Heat maps showing the distribution of population genetic variance (heritability) in the two-dimensional minor allele frequency (x-axis) and effect size (y-axis) space of causal variants.

Dark colors indicate that variants in this bin contribute a very small fraction of disease heritability, while white indicates a larger proportion (scale shown at bottom right).

As this figure shows, the effects of each parameter are quite intuitive. As the target size increases (going down rows in **Figure 2.8**), effect sizes decrease across the board because genetic risk is spread over a greater total number of causal variants. When the coupling to selection is high, variants under strong selection (very rare) have very large additive effects, and thus individuals who have a handful of such rare variants become affected with disease. The heritability, in this case, is concentrated among rare variants of large effect, and the upper left corner of **Figure 2.8** essentially shows an architecture that assumes T2D is a collection of Mendelian sub-types. When the coupling to selection is weak, heritability is concentrated among common variants of weak effect; the bottom right of **Figure 2.8** represents the common polygenic model advocated by the biometric school.

In summary, simple disease models with only two free parameters (target size and coupling to selection) are sufficient to generate a diverse set of genetic architectures with qualitative features that are consistent with prior expectation. These architectures have very different properties with respect to the contribution of rare vs. common variants and the range of causal variant effect sizes, and they include many of the most widely debated models of complex disease genetics.

References

1. Cowie, C., Rust, K., Byrd-Holt, D. & Gregg, E. Prevalence of Diabetes and High Risk for Population in 1988 – 2006. *Diabetes Care* **33**, 562-568 (2010).
2. Almgren, P. *et al.* Heritability and familiality of type 2 diabetes and related quantitative traits in the Botnia Study. *Diabetologia* **54**, 2811-9 (2011).
3. Zhu, Q. *et al.* A genome-wide comparison of the functional properties of rare and common genetic variants in humans. *American Journal of Human Genetics* **88**, 458-68 (2011).
4. Montgomery, S.B., Lappalainen, T., Gutierrez-Arcelus, M. & Dermitzakis, E.T. Rare and common regulatory variation in population-scale sequenced human genomes. *PLoS Genetics* **7**, e1002144 (2011).
5. Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476-82 (2011).
6. Ward, L.D. & Kellis, M. Evidence of Abundant Purifying Selection in Humans for Recently Acquired Regulatory Functions. *Science* **1675**, (2012).
7. McGuigan, K., Rowe, L. & Blows, M.W. Pleiotropy, apparent stabilizing selection and uncovering fitness optima. *Trends in Ecology & Evolution* **26**, 22-9 (2011).
8. Eyre-Walker, A. Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. *PNAS* **107**, 1752-6 (2010).
9. Tenesa, A. & Haley, C.S. The heritability of human disease: estimation, uses and abuses. *Nature Reviews Genetics* **14**, 139-49 (2013).
10. Park, J.-H. *et al.* Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. *PNAS* **108**, 18026-31 (2011).

Chapter 3

Application of the simulation framework to type 2 diabetes (T2D)

In the last two chapters, we simulated large, comprehensively genotyped and phenotyped human populations in which the full genetic architecture is known. A key aim of this framework, however, is to identify which architectures can and cannot be rejected for a particular disease on the basis of empirical data. In practice, the true genetic architecture of a trait cannot be directly read out; the only observable data are the results of genetic studies. Thus, in Chapter 3 we perform these studies (*in silico*) under each disease model. We then address our main question: which of the disease models described in Chapter 2 produce genetic study results that are compatible with observed data in genetic studies of type 2 diabetes (T2D)? Which genetic architectures are plausible for T2D?

Performing T2D genetic studies in simulated populations

Before we could directly evaluate each model, we needed to first define the set of genetic studies to conduct in simulated populations (to match studies conducted for T2D). We also needed to collate the empirical results of these studies (in European populations). These mainly included:

(a) epidemiological estimates of sibling relative risk (~1.8-3.4[1–3]);

(b) meta-analysis of linkage scans in ~4,200 affected sibling pairs (ASPs) with T2D (max LOD score 2.2[4]);

(c) *discovery* GWAS in 4,549 cases and 5,579 controls (DIAGRAMv1[5]; two genome-wide significant loci identified with p<5e-08);

(d) *replication* of the top (p<0.0001) signals from the discovery GWAS in an effective sample size of ~55K (~16 total loci met genome-wide significance[5]);

(e) *larger-scale meta-analysis* in 12,171 cases and 56,862 controls (DIAGRAMv3), followed by genotyping of top (p<0.005) signals on the Metabochip genotyping array in 34K cases and 115K controls[6,7] (39 genome-wide significant loci for T2D); and

(f) 'polygene score' logistic regression[8] using thousands of common marker effects learned in discovery GWAS (in aggregate, these explain 2.0-2.5% of test sample phenotypic variance, as measured by Nagelkerke's R^2).[9]

We performed each of the above studies under each of the disease models described in Chapter 2. At every step, we matched sample size and methodology to the empirical studies conducted for T2D. Importantly, each simulated genetic study was analyzed *without* knowledge of which variants were, in fact, causal (as would be the case in an actual study).

*Simulation of linkage and sibling measurements***.** From each simulated population, we sampled 10K unrelated cases and controls. Because we simulate nuclear families with multiple offspring in each generation (mean two offspring per mating), knowledge of sibling genotype and phenotype is available in each simulation. For each sampled case and control, we asked whether the individual's sibling was also affected with T2D. The fraction of cases' siblings who are affected divided by the fraction of controls' siblings who are affected yields the *sibling relative risk*. To perform affected sibling pair (ASP) linkage studies, we sampled 4200 sibling pairs in which both siblings are affected with T2D (matching the size of the largest European ASP meta-analysis for T2D[4]). SNP data provides a marker map that is significantly denser, but less polymorphic, than the microsatellite marker maps that were used in published studies. To model this, full sequence data was down-sampled across all causal loci; we included only variants with MAF>5% and pairwise LD (measured by r^2) < 0.2. The software package MERLIN was used to conduct non-parametric linkage analysis. The Z-scores resulting from such analyses are normally distributed; to generate LOD scores across 'background' non-causal loci, we randomly sampled 500 independent Z-scores from a normal distribution (representing a unique marker every ~5Mb of the human genome, similar to typical microsatellite map densities) and converted these to LOD scores using the relation: LOD = $Z^2/(2*ln(10))$. We recorded the genome-wide (across both causal and non-causal loci) maximum LOD score in each simulated study. Simulated models yielding a sibling relative risk of 1.8-3.5 (the

range observed across epidemiological studies of T2D) and no genome-wide LOD score greater than 3.0 (max LOD score observed for T2D was 2.2) were deemed consistent with empirical data.

Table 3.1: Simulation of multi-staged GWAS to match empirical T2D studies

Green rows represent stages of the below empirical studies at which simulated GWAS (discovery, replication, and large-scale meta-analysis) results are reported throughout the manuscript and figures.

Simulation of Zeggini et al, *Nature Genetics* 2008:
Discovery GWAS was simulated in 5K cases and 5K controls (matched to DIAGRAMv1). Top signals ($p<0.0001$) signals were carried forward for **replication** in an independent 20K cases and 35K controls.

Simulation of Morris et al *Nature Genetics* 2012:
A discovery GWAS was simulated in 16K cases and 16K controls (effective sample size matched to DIAGRAMv3). Top independent signals ($p<0.005$) were carried forward for genotyping (matched to Metabochip) in another 22K cases and 33K controls (effective sample size of ~52K), and combined with simulated DIAGRAMv3 for a **large-scale meta-analysis** in ~84K samples.

Publication	Stage of study design	Num. cases	Num. controls	Effective sample size*	
Zeggini et al *Nature Genetics* 2008	Stage 1 (DIAGRAMv1)	4,549	5,579	10K	**"Discovery (10K)"**
Zeggini et al *Nature Genetics* 2008	Stage 2 (DGI, FUSION, UKT2D)	10K	12K	22K	
Zeggini et al *Nature Genetics* 2008	Stage 3 (multiple cohorts)	14K	21K	33K	
Zeggini et al *Nature Genetics* 2008	Stage 1+2+3	29K	39K	65K	**"Replication (65K)"**
Morris et al *Nature Genetics* 2012	Stage 1 (DIAGRAMv3)	12K	57K	33K	
Morris et al *Nature Genetics* 2012	Stage 2 ('Metabochip')	22K	58K	52K	
Morris et al *Nature Genetics* 2012	Combined Stage 1+2	35K	115K	84K	**"Large-scale GWAS (85K)"**

* The sample size with equal numbers of cases and controls in which power would be equivalent to the power of the study conducted (calculated roughly as the quantity $2*N_{cases}*N_{controls})/(N_{cases}+N_{controls})$, calculated for each sub-cohort and summed across all cohorts included in meta-analysis). This value is significantly reduced as compared to the total number of cases and controls due to significant case-control imbalance (many more controls than cases) in several of the meta-analysis sub-cohorts.

Simulation of GWAS. We simulated several stages of GWAS to match studies conducted for T2D (**Table 3.1**). We simulated discovery phase GWAS for T2D (similar to DIAGRAM v1 stage 1) by sampling 4,549 cases with T2D and 5,579 controls from simulated populations under each disease model. To simulate commercial GWAS arrays, full-sequence data across all causal loci was down-sampled; we included only variants with MAF>5% and pairwise LD (measured by r^2) < 0.5. We

performed standard association analysis using the software PLINK. To model markers across background non-causal loci, we randomly sampled p-values between 0 and 1 to fill a total marker set of 2M SNPs (2.2M total SNPs were imputed, for comparison, in the DIAGRAMv1 study). We used the resulting distribution of genome-wide marker p-values to generate quantile-quantile plots and Manhattan plots for comparison to empirical data. We recorded the number of unique loci at which a marker p-value was < 5e-08. To simulate replication GWAS, we genotyped all markers from the discovery phase at which $p < 0.0001$ in 20K cases and 35K controls (effective sample size matched to that in DIAGRAMv1 replication), and performed association testing in this larger sample. The resulting p-values were used to determine the number of unique genome-wide significant loci predicted under each disease model after replication. Finally, we simulated large-scale GWAS in an effective sample size of ~35K total individuals, similar to DIAGRAMv3; we then simulated genotyping of all independent signals with $p<0.005$ on a genotyping array like Metabochip in an effective sample size of ~85K. When appropriate, sample sizes were corrected to account for imputation uncertainty, and p-values were adjusted to account for genomic-control corrections performed in empirical studies. The number of loci discovered at each stage of GWAS was compared to observed data for T2D from each published study. Simulated models yielding 1-4 genome-wide significant loci in discovery (N=10K; empirically 2 loci observed for T2D), 10-30 loci in replication (N=55K; empirically 16 loci observed for T2D), and 25-65 loci in large-scale meta-analysis (N=85K; empirically 39 loci observed for T2D) were deemed consistent with empirical data.

*Polygenic risk score analysis***.** Polygene 'score' analysis is a method by which to assess the aggregate predictive power of SNP alleles tested in a GWAS[8,9]. Following Stahl et al[9], we pruned common SNPs by their linkage disequilibrium, preferentially retaining the SNPs with lower discovery p-values to obtain a set of independent, maximally associated markers. We used the p-values and effect sizes from discovery GWAS to select subsets of SNPs reaching four different P_{GWAS} thresholds (0.001, 0.01, 0.1, and 0.5). For each SNP set, we summed the log-odds-weighted risk allele counts for each individual in an independent test sample of 2K cases and 3K controls to assign

each individual a polygene risk 'score'. We then tested these risk scores for association with case-control status using logistic regression. The predictive power of the polygene score was measured by Nagelkerke's R^2. Models yielding a Nagelkerke's R^2 between 0.01-0.04 for all P_{GWAS} thresholds were deemed consistent with data for T2D (empirically, Nagelkerke's R^2 was ~2-2.5% for all thresholds).

Evaluating the sensitivity of genetic study results to disease model parameters

We find, unsurprisingly, that the results of each study depend heavily on the underlying genetic architecture (**Figure 3.1**). The sibling relative risk (λ_s), for example, increases with increasing values of the selection parameter (τ); this occurs, intuitively, because under tightly coupled models (where $\tau = 1$), individual rare variants are relatively penetrant, and the chance that a sibling who shares a causal allele will also be affected with diabetes is high. We also observe that λ_s decreases as the target size (*T*) increases; the chance that two siblings share the same allele at all disease-causing sites decreases as the number of disease-causing sites grows.

We find that the results of ASP linkage scans are most sensitive to the disease target size. Even in sample sizes as large as 4200 ASPs, we find that linkage peaks are only consistently observed (regardless of the selection parameter) when the target size is small (*N*<30 loci). This is perhaps surprising, as it suggests that empirical studies were largely under-powered to test most models for a disease like T2D (with only modest heritability and high prevalence). Even under tightly coupled models (where some rare variants have large effects), linkage peaks (LOD>3) are not detected unless the target size is also small; when *N*=30 loci, linkage peaks are observed in >90% of simulation replicates, but when *N*=100 loci, only 20% of replicates produce a linkage peak. Thus, linkage data can only exclude a subset of oligogenic models for T2D, placing only limited bounds on the global genetic architecture of T2D.

The results of GWAS show interesting dependencies on both disease model parameters. Unsurprisingly (given that GWAS test only common variants), many more GWAS signals result under models where there is weak coupling to selection (and where the majority of disease

heritability is explained by common variants; see Chapter 2). When there is tight coupling to selection ($\tau = 1$), common causal variants have effects too weak to be detected in GWAS, and rare causal variants are poorly tagged by the common markers on GWAS arrays; as a result, very few GWAS signals are observed (regardless of target size). Under disease models with moderate or weak coupling to selection, both the number and strength of GWAS signals depend heavily on the target size, however. As the target size increases, genetic effects are spread out over more loci and the top GWAS signal decreases in strength. Initially (e.g. as the target size increases from *N*=30 to *N*=100 loci), the number of unique loci identified in GWAS increases, simply because there are exist a greater total number of causal loci. As the target size increases beyond 100 loci (*T*=250kb), however, the decreasing strength of signal at each locus causes the total number of genome-wide significant signals to decrease (**Figure 3.1, 3.2, 3.3**).

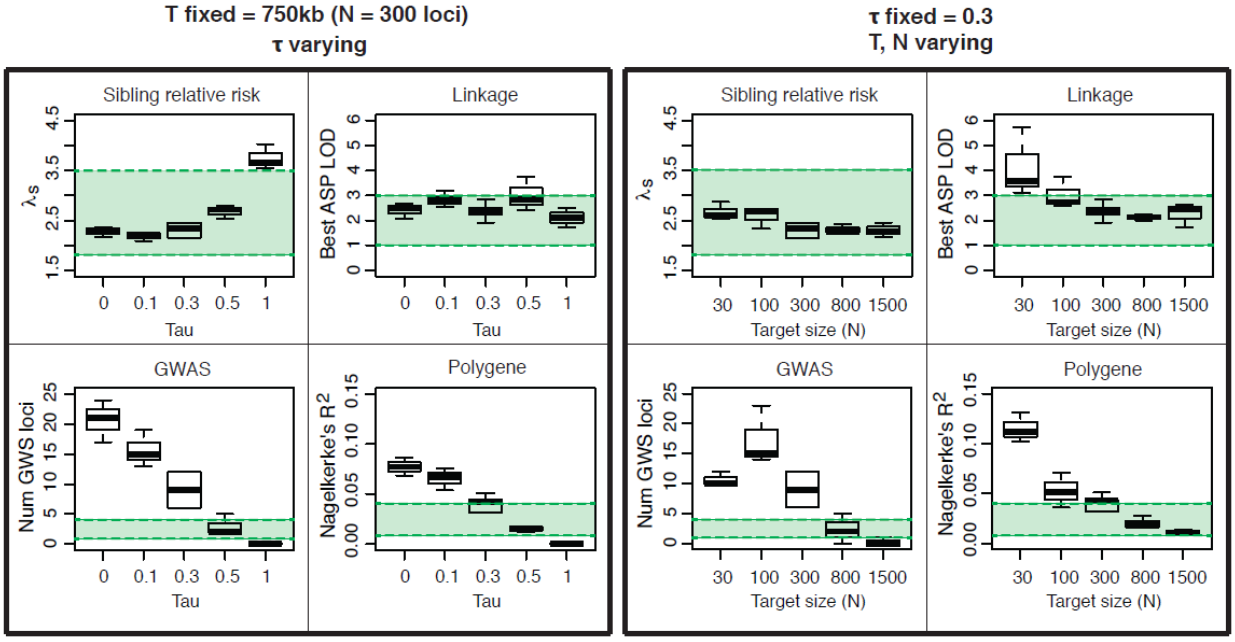


Figure 3.1: Sensitivity of genetic study results to disease model parameters.

Sensitivity of study results under models with *N* fixed at 300 loci and tau varying (**left box**) or tau fixed at 0.3 and *N* varying (**right box**). In each box, simulated data are shown (clockwise) for sibling relative risk, best genome-wide LOD score in an affected sibling pair (ASP) study of 4200 ASPs, number of genome-wide significant (p-value < 5*10^-8) loci detected in a GWAS of ~10K samples, and the Nagelkerke's R^2 value in a polygene score logistic regression in 5K samples using common variants with a discovery p-value < 0.01 ($P_{GWAS} = 0.01$). Green zones are centered (vertically) on empirically observed values for T2D, and represent the simulated values deemed consistent with empirical data.

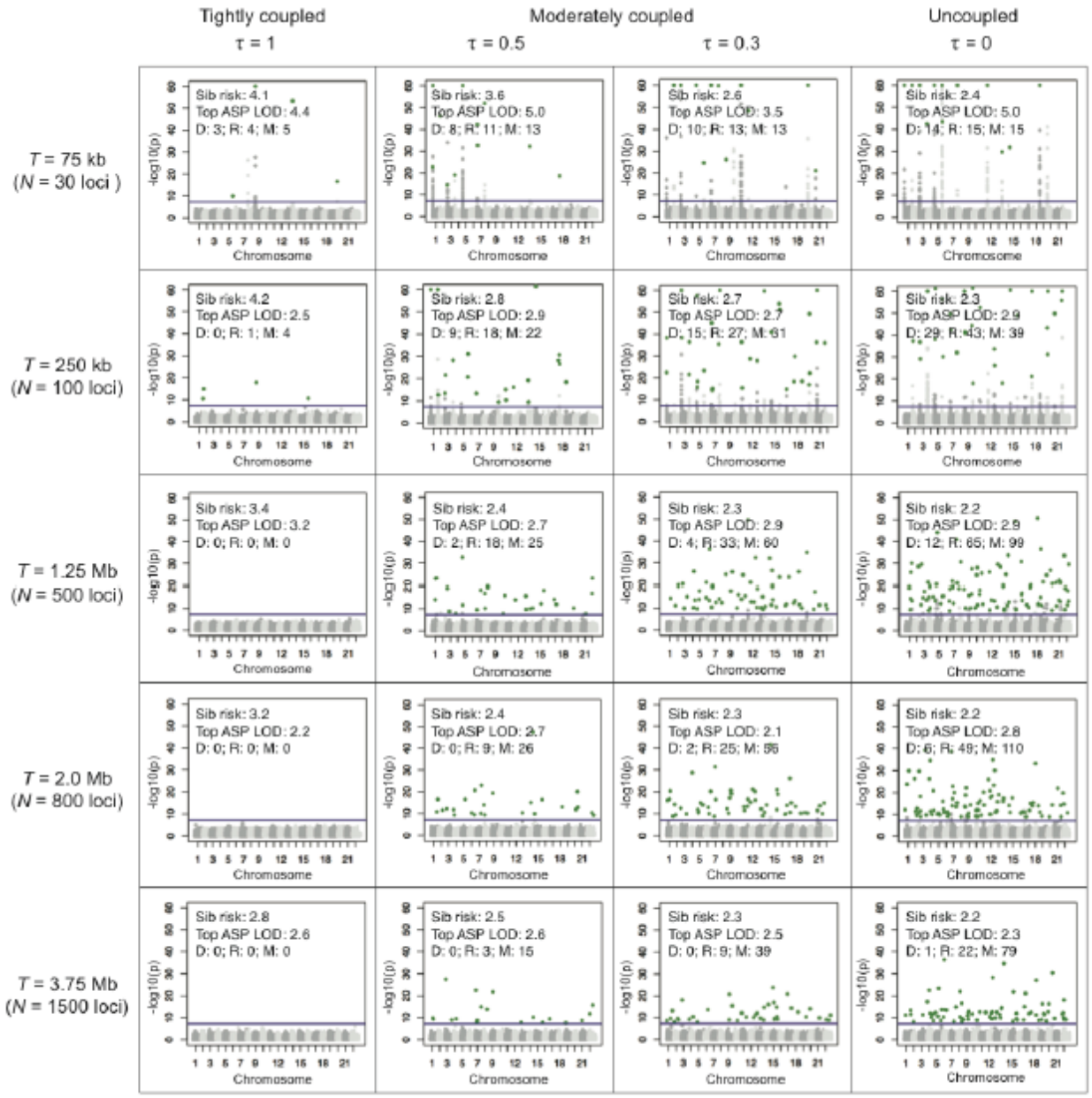


Figure 3.2: Results of sibling relative risk measurement, linkage, and GWAS under disease models.

Results are shown above across a range of simulated disease models, with varying selection parameter (columns) and target size (rows). Text indicates the median (across simulation replicates) sibling relative risk, median top genome-wide ASP LOD score, and the median number of genome-wide significant (p<5e-08) loci found after discovery GWAS (D; ~10K samples), replication of top signals in a larger study (R; ~55K samples, as in Zeggini et al 2008[5]), and larger-scale meta-analysis (M; ~80K effective sample size). The Manhattan plot represents data from the large-scale meta-analysis (simulated to match Morris et al, 2012[6,7]).

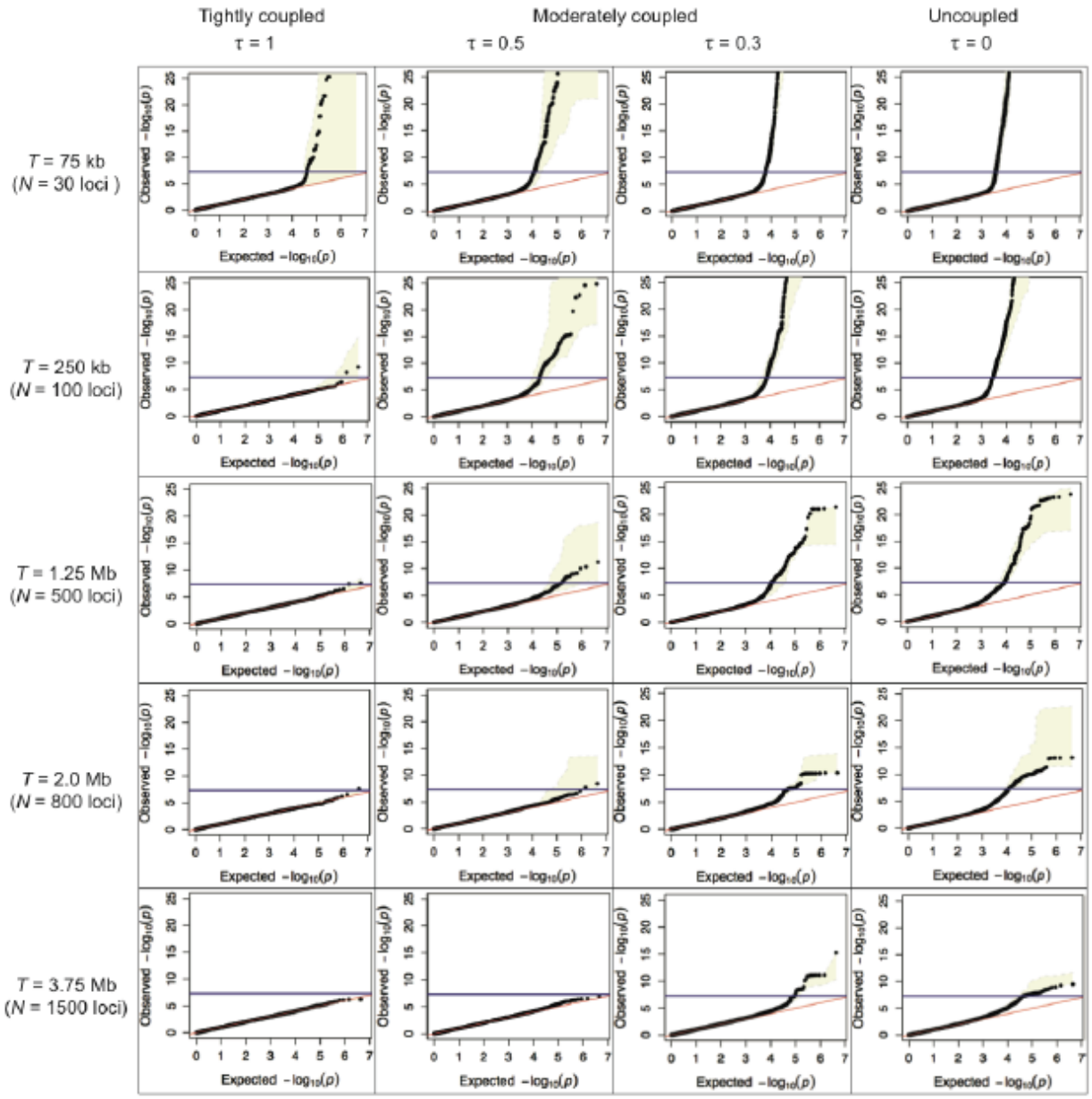


Figure 3.3: Simulated results of GWAS in ~10K samples under range of disease models.

Quantile-quantile plots above represent results of discovery GWAS (in ~5K cases and ~5K controls) under different disease models. Median across simulation replicates is shown above in black; beige shaded area represents range across replicates.

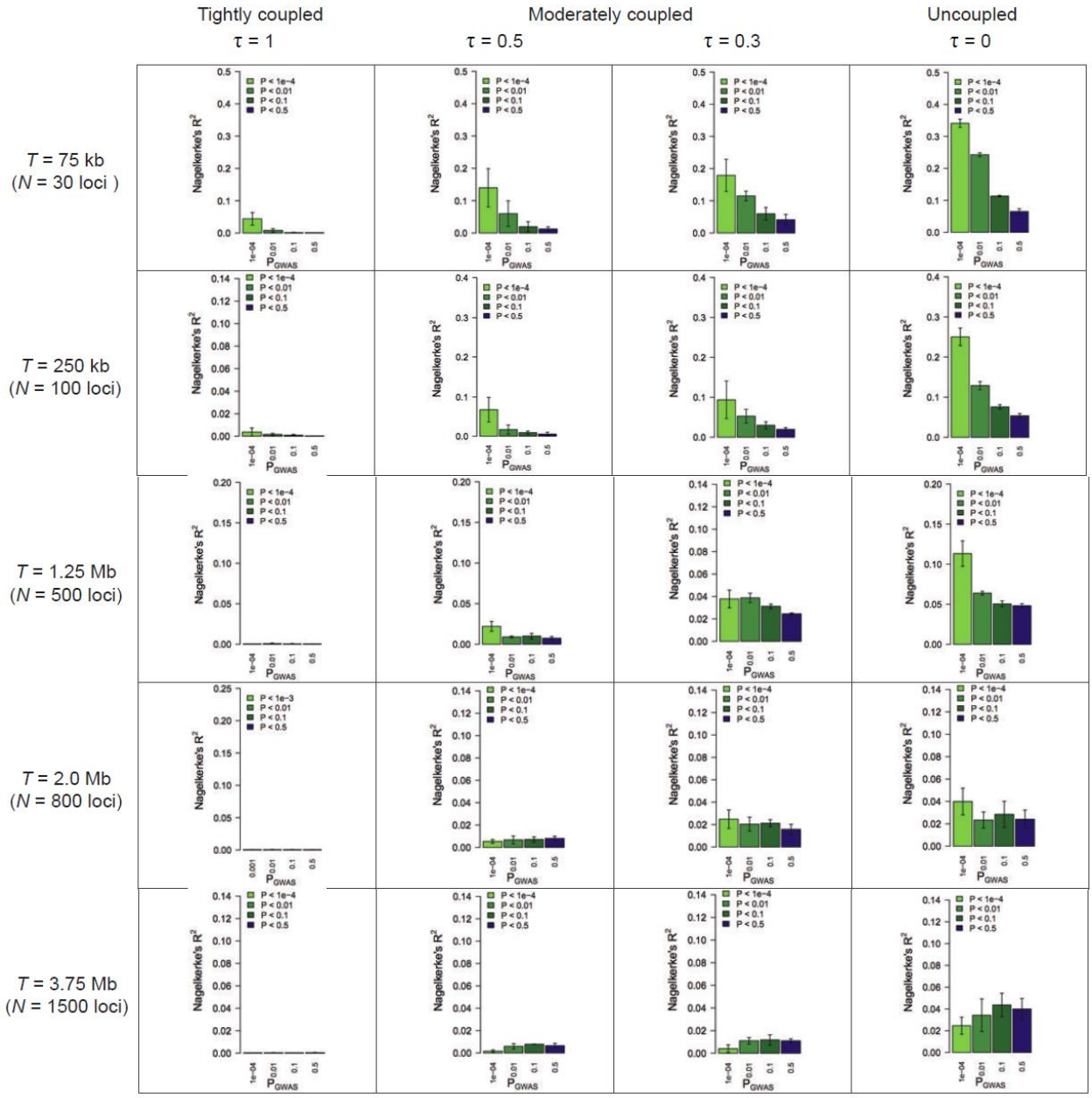


Figure 3.4: Simulated polygene score logistic regression under range of disease models.

Individuals were assigned polygene 'scores' using the log-odds-weighted sum of risk-allele counts, with odds ratios learned in a discovery GWAS of ~15K samples. Polygene R^2 represents Nagelkerke's R^2 in a logistic regression of phenotype vs. polygene score in an independent test sample of 2K cases and 3K controls (as in Stahl et al 2012[9]). Scores were computed using different p-value thresholds from the discovery GWAS. Error represents standard deviation across simulation replicates.

Finally, we measured the results of polygene score regression under each disease model. Given that this regression is performed using the full distribution of common variant signals observed

in GWAS, the R^2 metric quantifies the fraction of test sample phenotypic variance captured by *common marker* genotypes. We would expect, therefore, that this R^2 would be highest under uncoupled ($\tau = 0$) models where common variants explain the majority of genetic variance. Indeed, we find that the R^2 increases as the selection parameter decreases. Given a target size of 300 loci, for example, the R^2 ranges from < 0.001 when $\tau = 1$ to over 0.10 when $\tau = 0$. As the target size increases, the polygene R^2 value decreases as common variant effect sizes become weaker, and their estimates become noisier in discovery GWAS (**Figure 3.1, 3.4**).

Comparing simulated genetic study outcomes to empirical data for T2D

Having characterized the sensitivity of genetic study results to underlying properties of the disease models, we next evaluated each model for consistency with results for T2D. Evaluation of the full space of models is shown in **Figure 3.5**; detailed comparison for a few selected models (those highlighted in yellow in **Figure 3.5**) is shown in **Figure 3.6**.

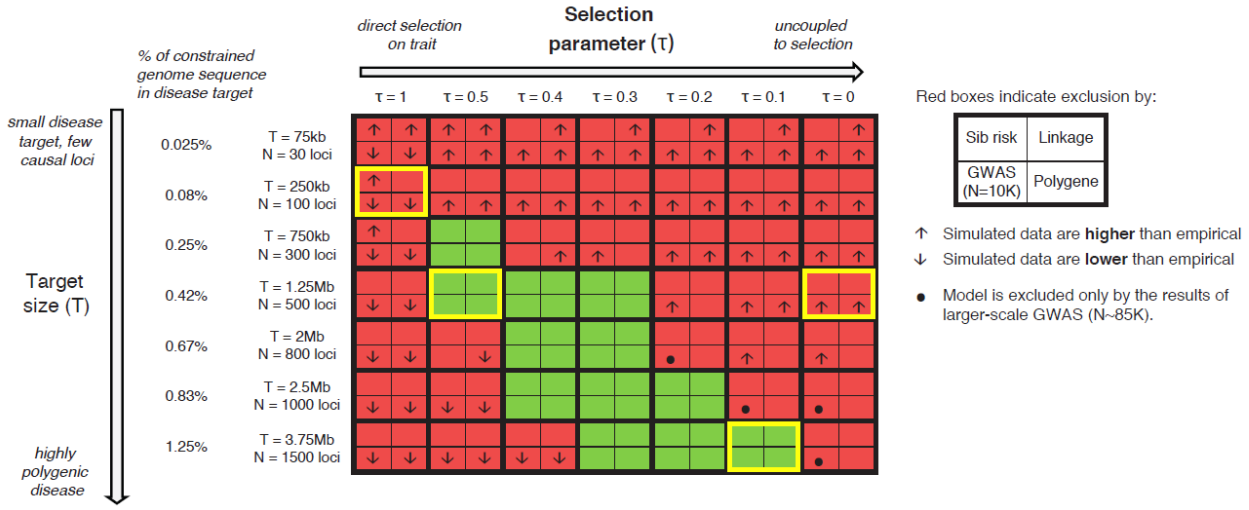


Figure 3.5: Evaluation of the full space of tested disease models against empirical data for T2D.

Space of disease models tested, each varying in target size (vertical axis) and selection coupling (horizontal axis). All models have fixed prevalence (8%) and heritability (45%), matching values observed for T2D. Each model produces results that are either inconsistent (red) or consistent (green) with empirical data for T2D. Inside red models, arrows indicate whether simulated results were too high or too low relative to empirical data. Dots in GWAS boxes indicate that the model is excluded by an excess of findings in large-scale (N~85K) GWAS (though results in 10K samples are consistent).

We first evaluated 'tightly coupled' models (τ=1) for consistency with observed data (left-most column of the grid in **Figure 3.5**). As seen in **Figure 3.1**, these models produce relatively high sibling risk (λ_s> 4) due to rare, high effect mutations shared by ASPs. The observation of linkage peaks (LOD>3.0) at small target sizes can exclude this subset of tightly coupled hypotheses, as no peaks were observed in empirical data for T2D. We find that GWAS results are sufficient, however, to exclude *all* 'tightly coupled' models, regardless of target size: under complete coupling, too few GWAS signals are observed even after large-scale follow-up (4-5 loci when *T*=250kb, for example, compared to 39 in empirical data; **Figure 3.6b**). Under tightly coupled models, polygene score regression is also less successful than empirically observed (R^2<0.5%, compared to ~2% for T2D). Thus, this class of disease models is globally inconsistent with data for T2D; it is unlikely that alleles under strong purifying selection have the largest effects on T2D, and it is thus unlikely that the T2D phenotype has been under direct evolutionary selection.

Next, we evaluated 'uncoupled' (τ=0) hypotheses (right portion of grid in **Figure 3.5**). These models produce modest risk to sibs (λ_s≈2), consistent with observed data. However, across a wide range of uncoupled models (up to *T*=3.75Mb, or *N*=1500 loci), an excess of GWAS findings is observed, as compared to empirical data. An example of such a model (τ =0, *T*=1.25Mb, or *N*=500 loci) is shown in **Figure 3.6d**; 11-19 GWAS loci are found in discovery (compared to 2 in empirical data), 61-71 loci after replication (16 empirically), and 99-102 loci in the large-scale GWAS followed by Metabochip genotyping (39 empirically). Under this uncoupled model, polygene score regression also explains a larger proportion of phenotypic variance than observed for T2D (R^2>10% at *p*<1e-4, compared to ~2% in empirical data, **Figure 3.6d**). While it is possible that uncoupled models with an even larger target size (*T>3.75Mb*, or *N*>1500 loci; not tested here) might be consistent with T2D data, these data suggest there is likely at least moderate coupling to selection for this phenotype. Biologically, we might interpret this to mean that the alleles contributing to T2D risk are, as a class, likely to influence phenotypes that are under some degree of negative selection; they may influence underlying metabolic traits, for example, which historically reduced evolutionary fitness.

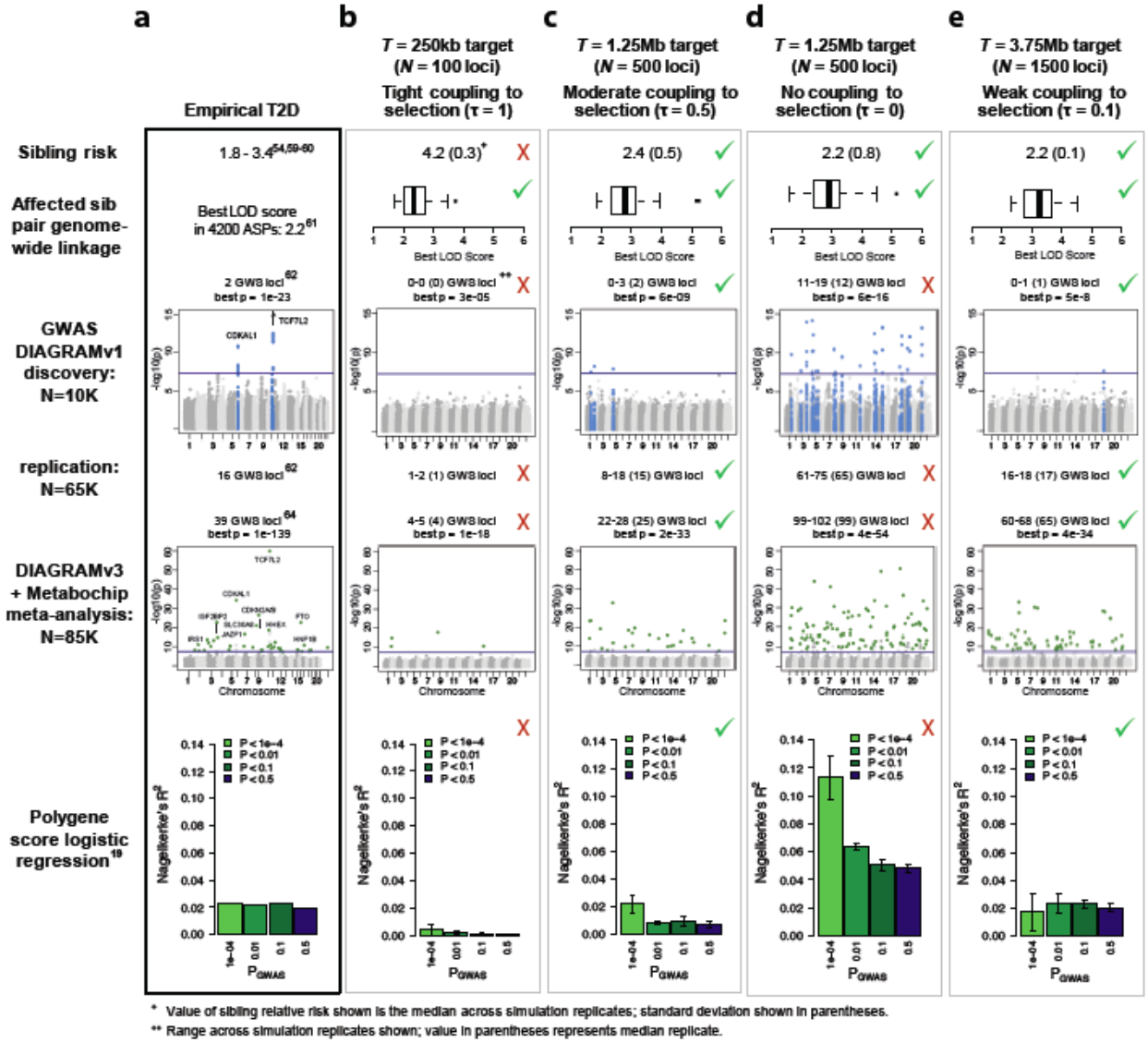


Figure 3.6: Simulated study results under representative disease models and comparison to T2D empirical data.

At left **(a)** are shown empirical genetic study results for type 2 diabetes (black outline). To right are shown simulated genetic study results for four different disease models. **b)** *T* = 250kb (*N* = 100 loci), τ =1 (tight coupling to selection); an 'extreme' rare variant model. **c)** *T* = 1.25Mb (*N* = 500 loci), τ =0.5 (moderate coupling to selection); an intermediate model. **d)** *T* = 1.25Mb (*N* = 500 loci), τ =0 (no coupling to selection); a 'common polygenic' model. **e)** *T* = 3.75Mb (*N* = 1500 loci), τ =0.1 (weak coupling to selection); a highly polygenic hybrid model. Red crosses indicate inconsistency with empirical data for T2D; green checks indicate consistency with empirical data. 'GWS loci' refers to the number of unique loci at which a variant isassociated to disease at genome-wide significance levels (p<5e-8).

While extreme models of genetic architecture are inconsistent with empirical data, a broad continuum of intermediate models remains consistent (green boxes in **Figure 3.5**). This class of

consistent models includes those with moderate coupling and smaller target sizes, as well as those with weak coupling and larger target sizes. Two examples are shown in **Figure 3.6c, e** ('moderate'; τ=0.5, *T*=1.25Mb, or *N*=500 loci) and ('weakly coupled'; τ=0.1, *T*=3.75Mb, or *N*=1500 loci). Predicted outcomes under both models are fully consistent with empirical data. However, these architectures have quite distinct properties: under the 'moderate' model, rare (MAF<5%) alleles explain ~80% of heritability, while under the 'weakly coupled' model, rare variants explain <25% of heritability. These data indicate that strong statements about global genetic architecture – at least statements based on only the findings of linkage and GWAS – are, as yet, premature because many disease models with widely varying properties remain consistent with empirical data for T2D.

As an aside, we note that it is even *harder* to constrain the locus-specific genetic architecture of T2D. Even if models assuming tight coupling to selection across *all* disease loci are inconsistent with empirical data for T2D, for example, it is still possible that individual T2D loci might harbor rare variants of large effect. We explored this possibility in **Figure 3.7**, where we further study the disease model with tight coupling to selection (τ=1) and a target size of *T*=250kb (*N*=100 loci). This architecture is globally inconsistent with empirical results for T2D (**Figure 3.6**), but we wondered whether a locus-specific architecture that arises under this model may still be plausible at individual T2D loci. Specifically, we asked whether synthetic associations[10] could arise under tightly coupled models, and whether linkage peaks would be expected in such cases. In **Figure 3.7**, we highlight two example simulated loci at which there exist rare variants of large effect, and where the common markers tagging the haplotypes on which these rare variants are present show (genome-wide significant) association to disease in a GWAS. Both loci are silent in a linkage scan, however, and thus cannot be excluded on the basis of negative linkage findings, and are entirely consistent with positive GWAS findings. These data underscore the challenge of constraining the locus-specific allelic architecture at complex disease loci. Chapter 4 addresses this problem by evaluating the performance of rare variant association methods at complex disease loci, and Chapters 5-6 directly test rare variant genetic models at T2D loci using empirical sequencing data across these loci.

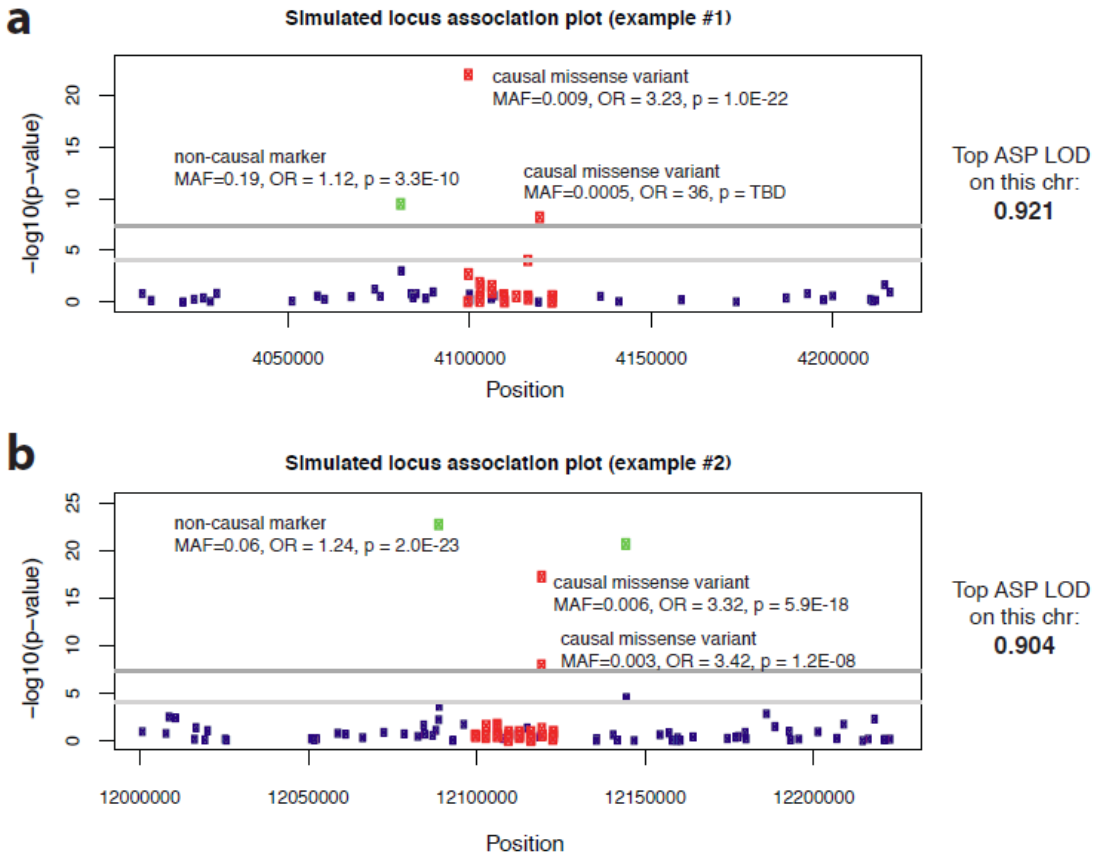


Figure 3.7: Synthetic associations could be observed under tightly coupled models and cannot be excluded by empirical linkage results.

Shown above are two example loci simulated under a disease model with tight coupling to selection (tau=1) and a target size of 250kb (N=100 loci). Blue points represent common (MAF>5%) markers used for the GWAS; green points are markers achieving genome-wide significant association; red points are causal variant at the locus (they would not have been typed in a GWAS, but if they were subsequently assayed their signals would be as shown). Horizontal lines indicate p=1e-4 and p=5e-8.

What will it take to know? Predicting the results of future and ongoing genetic studies of T2D

Ongoing studies are now using (a) exome and whole-genome sequencing and (b) genotyping via an exome array to study rare and intermediate frequency variants in modest (thousands) and large (tens of thousands) samples, respectively. In coming years, it is predicted that sequencing will be performed in hundreds of thousands or even millions of people. We thus asked: to what extent will these ongoing and future studies further constrain T2D genetic architecture?

We simulated high-coverage, whole-genome sequencing of 3K and 10K individuals (sample sizes similar to those of studies being performed by the Go-T2D and T2D-GENES Projects, respectively), as well as a study in which a large proportion of rare coding variants are genotyped in 20K cases and 35K controls (also similar to ongoing Exome Chip studies for T2D). In each study, we simulated both single variant association as well as rare variant gene-based burden testing. To project studies that might be done in coming years, such as in the UK Biobank, we simulated complete genome sequencing of an unselected population cohort of 250K individuals (20K cases, 230K controls).

We then asked: at what point will disease models that are currently consistent with all available data diverge in these future studies? As examples, we focused on the two consistent models depicted in **Figure 3.6c** ('moderate') and **Figure 3.6e** ('weakly coupled'). For both models, whole-genome sequencing in 3K individuals discovers few signals not previously detected by GWAS (**Figure 3.8**). In 10K samples, the models diverge slightly: ~15 novel loci (representing ~6% of heritability) are predicted under the 'moderate' model, whereas ~5 loci (representing <1% of heritability) are predicted under the 'weakly coupled' model. The most significant constraint, however, is predicted to come from large exome array studies: we predict ~80 novel loci under the 'moderate' model (bringing cumulative heritability explained to ~50%), but only ~10 loci under the 'weakly coupled' model (and ~15% of heritability explained). Thus, at least one of these models will likely be inconsistent with the results of studies already planned for T2D.

As sample size is expanded to 250K unselected individuals, these models diverge further. In both cases, substantial discovery is predicted, but the total fraction of heritability explained, as well as the frequency distribution of identified causal variants, differs. Under the 'moderate' model, over half (~265 out of 500) of all disease loci would be discovered, and would collectively explain ~75% of T2D heritability. At a majority of these loci, the most disease-associated variant would be rare (MAF<2%). Under the 'weakly coupled' model, a much larger fraction of disease loci would remain undetected (due to the individually small effect sizes of very many causal variants), and a smaller

proportion of total heritability (~48%) would be explained. However, the most associated variant at virtually all of these loci would be common (MAF>2%), and thus likely discoverable by GWAS alone of comparable sample size, without need for complete sequencing.

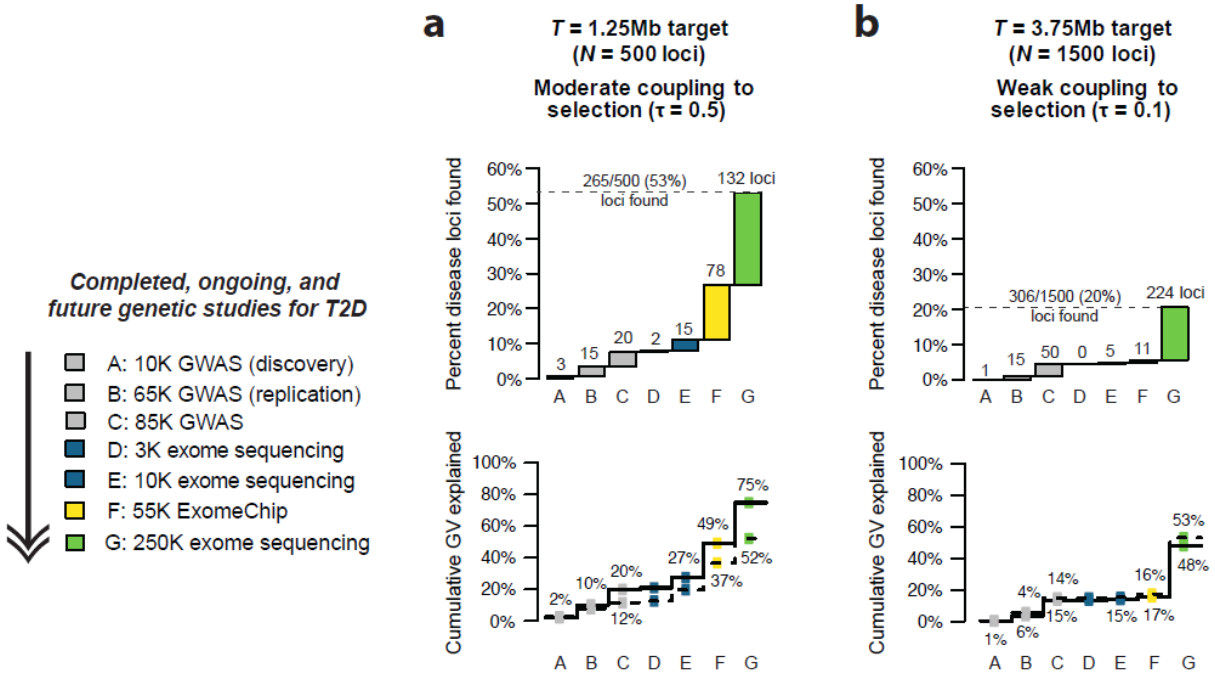


Figure 3.8: Prediction of ongoing sequencing and large-scale genotyping studies for T2D under different disease models that are currently consistent with empirical data.

Predictions under the two consistent disease models from **Figure 3.6** are shown here: **(a)** a model with 'moderate' coupling to selection and a target size of *T*=1.25Mb (*N*=500 causal loci), and **(b)** a 'weakly coupled' model with a target size of *T*=3.75Mb (*N*=1500 causal loci). Top charts show cumulative fraction of disease loci discovered by each study design: A = Discovery GWAS in 10K samples, followed by B = Replication genotying of top signals in 55K independent samples (as in Zeggini et al 2008); C = large-scale GWAS with discovery in an effective sample size of ~30K, followed by genotyping all independent signals with p<0.005 (as on Metabochip) to yield a total effective sample size of ~85K (as in Morris et al 2012); D = exome sequencing in 3K samples; E = exome sequencing in 10K samples); F = genotyping in 20K cases and 35 controls all rare variants (exome-wide) seen >= 2x in 5K controls (similar to Exome Chip); G = exome sequencing in 20K cases and 230K controls (a 250K unselected population cohort with T2D prevalence 8%). Labels above bars indicate predicted number of novel loci (e.g. not found in the previous studies) discovered at each step. Bottom charts show cumulative fraction of population genetic variance (heritability) explained by loci uncovered in each study. Solid line indicates true variance explained by those loci; dotted line represents fraction estimated using frequencies and odds ratios (estimated in the study) of the most associated single variants at each locus.

Thus, ongoing sequencing and genotyping studies (and the extent to which they are successful) will likely place substantial bounds on T2D genetic architecture. However, studies of

hundreds of thousands of individuals will likely be required to discover many if not most of the genes underlying T2D, and even then a substantial fraction of heritability (and of causal loci) may remain undiscovered. This is not meant as nihilistic – already much has been learned about the genetic basis of T2D, and our study suggests that in coming years a great deal more will be discovered, including further constraints on models of genetic architecture. The challenge of localizing disease heritability may simply be the expected outcome for a population genetic process which results in many causal alleles, strong and weak, that are both common and rare.

Limitations of this approach and future steps

This study (the framework described in Chapters 1-3) has a number of limitations. Although only two model parameters were sufficient to generate diverse architectures, more parameters could be included. For example, causal variants were simulated only at regions under purifying selection; alternate models could be explored in which neutrally evolving alleles also have effects on disease. Positive selection was not simulated, and derived alleles were only modeled as increasing disease risk (though interestingly, this does not preclude the occurrence of significantly associated markers of protective effect; data not shown here).

Additionally, locus structure in our study was uniform; we considered only causal loci modeled as protein-coding genes and did not model the structural properties of non-protein-coding functional elements. Adding skew in the distribution of length, overall phenotypic contribution, and coupling to selection across disease loci could also produce more varied models. Finally, non-additive inheritance, epistasis, or gene-environment interactions were not modeled. In future work, if the outcomes of many types of genetic studies (such as those directly simulated here) in human populations could be accurately predicted using analytical solutions, then an inferential approach may enable efficient traversal across disease models defined by many more variable parameters.

References

1. Almgren, P. *et al.* Heritability and familiality of type 2 diabetes and related quantitative traits in the Botnia Study. *Diabetologia* **54**, 2811-9 (2011).
2. Lyssenko, V. *et al.* Predictors of and longitudinal changes in insulin sensitivity and secretion preceding onset of type 2 diabetes. *Diabetes* **54**, 166-74 (2005).
3. Weijnen, C.F., Rich, S.S., Meigs, J.B., Krolewski, a S. & Warram, J.H. Risk of diabetes in siblings of index cases with Type 2 diabetes: implications for genetic studies. *Diabetic Medicine* **19**, 41-50 (2002).
4. Guan, W., Pluzhnikov, A., Cox, N.J. & Boehnke, M. Meta-analysis of 23 type 2 diabetes linkage studies from the International Type 2 Diabetes Linkage Analysis Consortium. *Human Heredity* **66**, 35-49 (2008).
5. Zeggini, E. *et al.* Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nature Genetics* **40**, 638-45 (2008).
6. Voight, B.F. *et al.* The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS Genetics* **8**, e1002793 (2012).
7. Morris, A.P. *et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature Genetics* **44**, (2012).
8. Purcell, S.M. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748-52 (2009).
9. Stahl, E. *et al.* Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. *Nature Genetics* 1-9 (2012).doi:10.1038/ng.2232
10. Dickson, S.P., Wang, K., Krantz, I., Hakonarson, H. & Goldstein, D.B. Rare variants create synthetic genome-wide associations. *PLoS Biology* **8**, e1000294 (2010).

Chapter 4

The power of gene-based association methods under different locus architectures

The advent of next-generation sequencing studies in large case-control cohorts are now enabling insights into the contribution of rare variation to complex trait heritability for the first time. These studies, however, probe an almost limitless number of heterogeneous hypotheses about allelic architecture. In this chapter (for which work was performed in close collaboration with Loukas Moutsianas at Oxford), we simulate a subset of the complex trait genetic architectures described in Chapter 2 at human disease loci, and evaluate the extent to which currently available gene-based association methods can identify such signals in sample sizes comparable to those of ongoing re-sequencing studies. We assess the impact of locus architecture, effect size, and functional variant filters on the power of each method at stringent levels of significance. By evaluating all tests together at loci simulated under a range of continuous frequency-effect size distributions, we are able to characterize each method's specific success and failure modes, and describe genetic hypotheses for which particular methods may be optimally powered.

Background on gene-based rare variant association methods

Single-variant association tests have limited ability to interrogate the role of rare (defined here as MAF<1%) genetic variation in disease; power to detect a variant with MAF 0.5% and RR 2.5 in 3K case-control samples at α=5e-8, for example, is less than 5%.[1] Variants that are private to individuals, as some deleterious mutations are hypothesized to be, present greater challenges yet, as their association cannot be individually detected. As a result, numerous statistical methods have been developed in recent years to test groups of rare variants in aggregate for association to disease.[2–4]

A handful of targeted re-sequencing experiments have recently identified rare variants which modulate risk for common, complex diseases. Examples include variants in *NOD2* for Crohn's

disease (4 variants with MAF 0.1-0.8%, ORs 1.4-4.0)[5], *PCSK9* for coronary heart disease (2 variants with MAF 0.8 and 1.8%, OR ~0.1)[6], *LPL* for hypertriglyceridemia (154 missense variants with MAF<1% in cases, detected using the T1 method)[7] and *MTNR1B* for type 2 diabetes (13 functionally-screened variants with MAF<0.1%, collective OR ~5.5, detected using the KBAC method)[8]. Each of these disease loci is characterized by different numbers, frequencies, and effect sizes of rare and low frequency variants, but in each of these selected examples, the estimated proportion of phenotypic variance explained per locus is ~0.5-1.5%. It is also worth noting that in most of the above examples, the levels of statistical significance attained were modest: sufficient for a candidate gene study, but less informative in the setting of larger-scale sequencing studies.

As such larger-scale studies are conducted (e.g. exome-wide or genome-wide) in thousands of samples, a number of questions emerge. Are loci similar to *LPL* or *MTNR1B* scattered across the genome? If so, what is the power of different gene-based methods to detect them? What effect sizes are required for studies of a given sample size to be well-powered, and what significance thresholds are appropriate? In order to interpret the results of gene-based association methods in sequencing studies, it is critical to quantify the power of each method to detect signals under hypothesized locus architectures.

Although the introduction of each novel gene-based association test has typically been accompanied by evaluation of the method's performance alongside some alternatives, each such analysis compared different subsets of tests, made different assumptions about locus architecture and study design, and employed different simulation approaches. A few studies have documented the relative power of different methods[9–11], but they evaluated only a few methods (not including those most recently published), did so in small sample sizes, simulated *ad hoc* locus architectures (e.g., with fixed numbers of causal variants) that may not be representative of complex diseases, and considered only nominal levels of significance ($\alpha>0.01$). Thus, it is as yet unclear which gene-based methods investigators should use to test specific genetic models of disease.

Simulation of diverse genetic architectures at human haplotypes

We developed a new simulation approach – informed by the simulations conducted in Chapters 1-3, but nonetheless distinct – to address these questions. Specifically, we wanted to perform simulations at real human gene loci, where both the locus structure and haplotype diversity would reflect empirical reality.

To this end, we adapted the software HAPGEN2[12] to simulate variation across the full SFS in thousands of samples. We started with phased haplotypes from 379 European samples (1000G Project Phase 1, release 3).[13] To expand this reference panel to a larger number of samples, we applied a staged, iterative approach which preserves linkage disequilibrium structure between relatively common variants while introducing new low-frequency variants upon the original haplotypes. We then down-sampled variants to match the empirical SFS recently observed at protein-coding genes in over 12K samples[14] (**Figure 4.1**). We selected 24 human protein-coding loci of average coding length on chr10 (**Figure 4.1a, Table 4.1**) at which to perform simulations.

As in Chapters 2-3, we modeled the complex disease type 2 diabetes (prevalence 8%). We introduced phenotypic effects (odds ratios per variant) by sampling multiple causal variants per locus from six different joint distributions of causal variant frequencies and effect sizes (**Table 4.2**, **Figure 4.2**). These distributions were learned from forward simulations of global genetic architecture under different disease models (Chapter 2); importantly, this ensured a principled distribution of causal variant frequencies and effect sizes (consistent with T2D prevalence and heritability) rather than an *ad hoc* distribution. The three main architectures assumed very strong (AR1: τ=1), moderate (AR2: τ=0.5), or weak (AR3: τ=0) purifying selection against causal alleles. As seen in Chapter 2, AR1 results in a sharp inverse correlation between variant frequency and effect size, AR2 produces modest correlation, and AR3 is characterized by rare and common alleles that have similar effects on phenotype. We simulated two additional architectures, AR4 and AR5, which are variations of AR1 and AR2, respectively, where *only* rare (MAF<1%) variants at a locus contribute to phenotype. Finally, AR6 assumes a frequency-effect size map identical to AR2, but assigns a 50%-50% mixture

of risk and protective effects; this represents the hypothesis that some variants in a gene increase disease risk, while other variants in the same gene can have a protective effect.

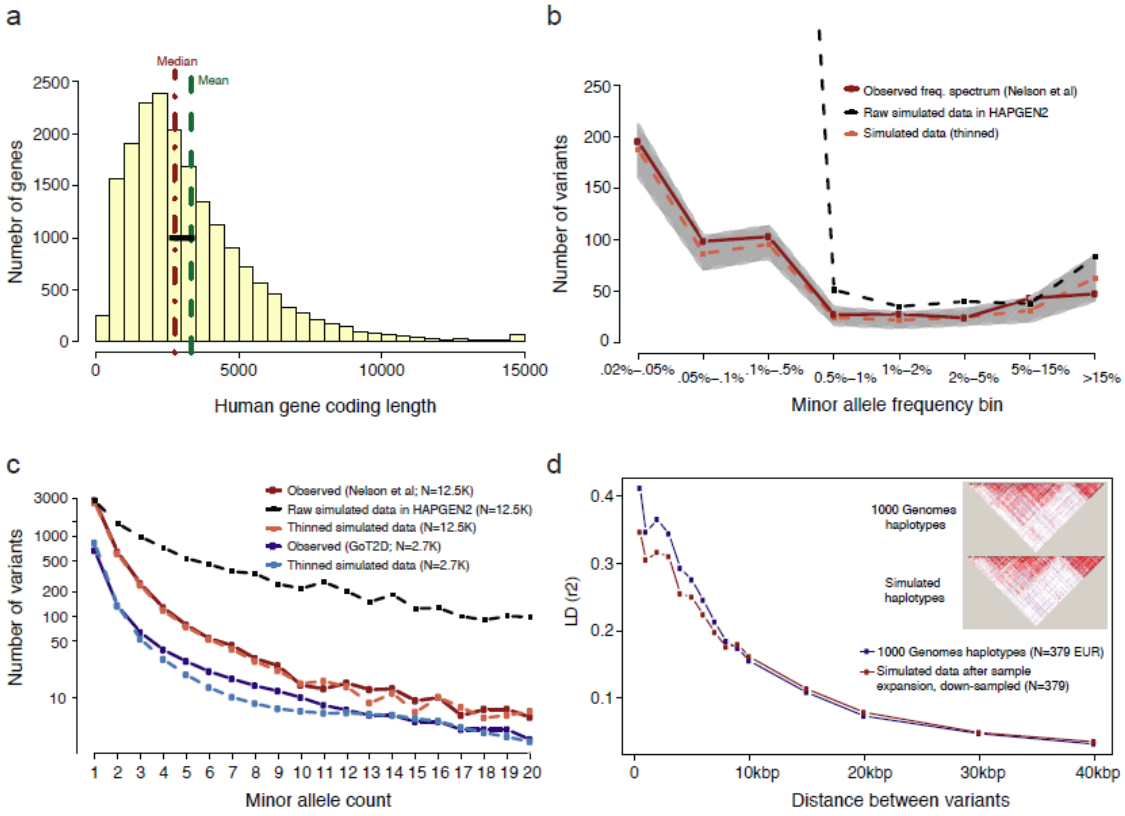


Figure 4.1: Generation of simulated genotype data at human loci in large sample sizes with HAPGEN2

Haplotypes were simulated at 'average' human protein-coding genes drawn from the center of the distribution of RefSeq gene total exon length **(a)**. Vertical dotted lines in red and green indicate the median and mean values of exon length, respectively. Blue points represent the 24 genes selected for simulation. **(b,c)** Simulated site frequency spectrum, as compared to observed human data. Data were simulated via staged expansion of 1000 Genomes Project haplotypes using the HAPGEN2 software; the mutation parameter was fit to match the site frequency spectrum of protein-coding variation observed in exome sequencing studies, e.g. as reported Nelson et al 2012. Raw simulated data from HAPGEN2 in large sample sizes producedan excess of rare sites; these were down-sampled to match observed data. The grey area in **(c)** represents the [5%,95%] interval across all simulated genes, obtained using bootstrapping. The site frequency spectrum of simulated data in a smaller sample size (N=2.7K) also matched an independent set of observed exome sequencing data from the GoT2D consortium **(c)**. Haplotype structure, as measured by linkage disequilibrium between variants, was also preserved in the simulated data after sample expansion **(d)**. Inset shows example simulations at the GATA3 gene locus.

Table 4.1: Human protein-coding loci at which simulations were performed

24 genes on chr10 were selected from the center of the distribution of human gene coding length (**Figure 4.1**). Below are the genomic locations of these regions; HAPGEN2 simulations were performed across the full length of each transcript; causal variants were selected from exonic regions only, and burden testing was also performed on variants (causal and non-causal) within the exonic regions only.

CHR	START (hg19)	END (hg19)	UCSC_Transcript	GENE	Transcript Length (kbp)	Refseq ID	STRAND	Total Exon Length (kbp)
chr10	104,221,169	104,236,802	uc001kvt.3	TMEM180	15633	NM_024789	+	2834
chr10	105,036,919	105,050,108	uc001kws.3	INA	13189	NM_032727	+	3231
chr10	1,085,963	1,091,076	uc001ifz.3	IDI1	5113	NM_004508	-	2843
chr10	115,999,017	116,051,737	uc001lbl.1	VWA2	52720	NM_198496	+	3330
chr10	116,697,951	116,737,439	uc001lcd.3	TRUB1	39488	NM_139169	+	3412
chr10	119,301,955	119,309,057	uc001ldh.4	EMX2	7102	NM_004098	+	2896
chr10	126,490,353	126,525,239	uc001lib.4	FAM175B	34886	NM_032182	+	2992
chr10	15,117,473	15,130,775	uc001inv.3	ACBD7	13302	NM_001039844	-	3370
chr10	23,216,953	23,326,514	uc001irm.4	ARMC3	109561	NM_173081	+	2808
chr10	26,505,235	26,593,491	uc001isq.2	GAD2	88256	NM_000818	+	2824
chr10	30,722,949	30,750,762	uc001ivi.2	MAP3K8	27813	NM_005204	+	3013
chr10	35,415,768	35,501,886	uc001iya.3	CREM	86118	NM_183013	+	2714
chr10	5,454,517	5,500,426	uc001iia.3	NET1	45909	NM_001047160	+	3398
chr10	6,052,656	6,104,333	uc001iiz.2	IL2RA	51677	NM_000417	-	3216
chr10	62,538,088	62,547,988	uc010qii.2	CDK1	9900	NM_001170406	+	2915
chr10	64,571,755	64,576,126	uc001jmi.3	EGR2	4371	NM_000399	-	2979
chr10	70,587,293	70,652,816	uc001joq.3	STOX1	65523	NM_001130162	+	3381
chr10	72,058,728	72,141,414	uc001jqx.1	LRRC20	82686	NM_207119	-	3158
chr10	73,576,054	73,611,082	uc001jsm.3	PSAP	35028	NM_002778	-	2822
chr10	75,670,861	75,677,258	uc010qkw.2	PLAU	6397	NM_001145031	+	2662
chr10	8,096,666	8,117,164	uc001ijz.3	GATA3	20498	NM_001002295	+	3070
chr10	85,980,248	85,985,284	uc010qmc.2	LRIT2	5036	NM_001017924	-	3118
chr10	89,512,874	89,577,917	uc001kez.1	ATAD1	65043	NM_032810	-	3034
chr10	96,305,573	96,361,856	uc009xuo.3	HELLS	56283	NM_018063	+	3237

Table 4.2: Locus architectures modeled at simulated human disease loci

The below range of locus architectures were modeled at simulated loci; variant effect sizes were sampled from joint frequency-effect size distributions learned from forward population genetic simulations (described in Chapter 2). The architectures were chosen to reflect a range of different rare variant contributions and effect sizes. At each locus, the total number of causal variants depended on the effect sizes sampled, as loci were modeled to explain a fixed proportion of liability-scale phenotypic variance underlying a complex trait with 8% prevalence.

Architecture	Direction of effects	Selection parameter (from Chapter 2)	Description
AR1	All deleterious	tau = 1	Rare variants (MAF<1%) explain >90% of heritability and have large effects relative to common variants
AR2	All deleterious	tau = 0.5	Rare variants explain ~50% of heritability and have moderate effect sizes relative to common variants
AR3	All deleterious	tau = 0	Rare variants explain <15% of heritability and have additive effects comparable to common variants
AR4	All deleterious	tau = 1, all causal variants have MAF < 1%	Same effect size distribution as AR1; but only rare (MAF<1%) are causal at the simulated locus
AR5	All deleterious	tau = 0.5, all causal variants have MAF < 1%	Same effect size distribution as AR2; but only rare (MAF<1%) are causal at the simulated locus
AR6	50% deleterious, 50% protective	tau = 0.5	Same effect size distribution as AR2; but 50% of variants have protective effects on disease

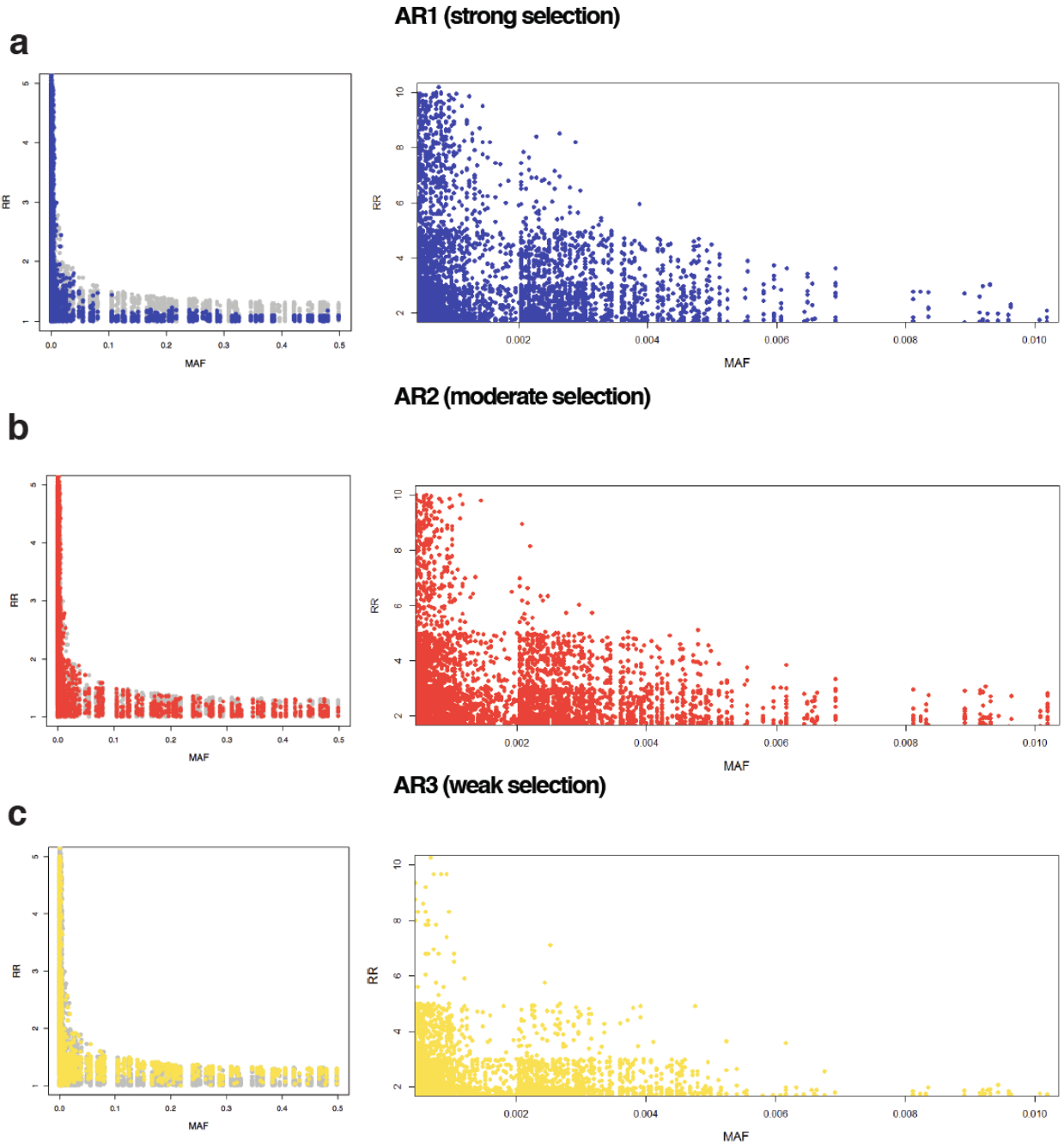


Figure 4.2: Variant frequency-effect size distributions under each simulated architecture

The below frequency-RR distributions were learned the simulations of global genetic architecture described in Chapter 2. AR1 assumes strong coupling to purifying selection; that is, variants under selection (more likely rare) have larger effects on disease. AR3 assumes no coupling to selection, and thus effect sizes are more uniform across the frequency spectrum.

Evaluation of gene-based rare variant association methods under simulated architectures

We evaluated a representative set of 10 gene-based association methods (BURDEN[15], C-ALPHA[16], CMC[17], FRQWGT[15], KBAC[18], MADSEN[19], SKAT[20], SKAT-O[21], UNIQ[15], and VT[22]; see **Table 4.3**) on these simulated datasets. The tests we applied can be broadly categorized as unidirectional 'burden' tests, bidirectional variance-component tests (SKAT, CALPHA), and a unified, linear combination of these two classes (SKAT-O). The unidirectional tests can be further sub-divided into collapsing regression methods (CMC), weighted sum methods (FRQWT, KBAC, MADSEN, VT), and permutation-based summary count methods (BURDEN, UNIQ). We selected this set of tests in part because they are readily available in the most recent releases of the widely-used software packages PLINK/Seq[15] and EPACTS[23]. Before further evaluation, we confirmed that all tests are well-calibrated under (null) datasets where no effects are assigned to any variants (**Figure 4.3**).

Table 4.3: Published gene-based rare variant association methods evaluated in this study

Method name	Citation	Software	Description
Unidirectional rare variant gene-based tests			
Collapsing methods			
Combined Multivariate and Collapsing (CMC)	Liu & Leal, PLoS Comp. Bio. 2008	EPACTS	All rare variants collapsed into a single variant; individual dosage for the collapsed 'variant' is regressed against phenotype.
Weighted and unweighted sum methods			
Variable threshold (VT)	Price et al, AJHG. 2010	PLINK-Seq	Sum of rare allele count in cases vs. controls; allele frequency threshold for inclusion is varied to maximize test statistic.
Weighted Sum Statistic (WSS; FRQWT)	Madsen & Browning, PLoS Gen. 2009	PLINK-Seq	Permutation-based test comparing inverse-frequency-weighted rare variant counts per individual in cases vs. controls.
Weighted Sum Method (Madsen)	Madsen & Browning, PLoS Gen. 2009	EPACTS	Wilcoxon Rank Sum test between phenotypes and inverse-frequency-weighted rare variant scores.
Kernel-Based Adaptive Cluster (KBAC)	Liu & Leal, PLoS Gen. 2010	PLINK-Seq	Variant weights are determined adpatively, and are based on observed effect sizes; individuals scored by weighted sum of allele counts.
Summary case:control count methods			
BURDEN method	Purcell (PLINK-Seq)	PLINK-Seq	Permutation-based test comparing raw allele counts in cases vs. controls.
UNIQ test	Purcell (PLINK-Seq)	PLINK-Seq	Simple count of total case-unique rare alleles; permutations to assess significance.
Bi-directional variance-component gene-based tests			
C-ALPHA	Neale et al, PLoS Gen. 2011	PLINK-Seq	Detects deviation of observed case:control variant counts from expected binomial distribution.
Sequence Kernel Association Test (SKAT)	Wu et al, AJHG 2011	EPACTS	Generalized form of C-ALPHA with variants weighted by allele frequency.
Linear combination of unidirectional and variance-component tests			
SKAT-O ('Optimal' SKAT)	Lee et al, AJHG. 2012	EPACTS	Adaptive linear combination of unidirectional burden test and variance-component SKAT test.

Each test was run on all exonic variants with MAF<1% (both causal and non-causal). The power of each method to detect a locus explaining 1% of the variance in disease liability[24,25] in 1500

cases and 1500 controls (sample size comparable to that of ongoing sequencing studies) is shown as a function of significance threshold (α) and architecture in **Figure 4.4**.

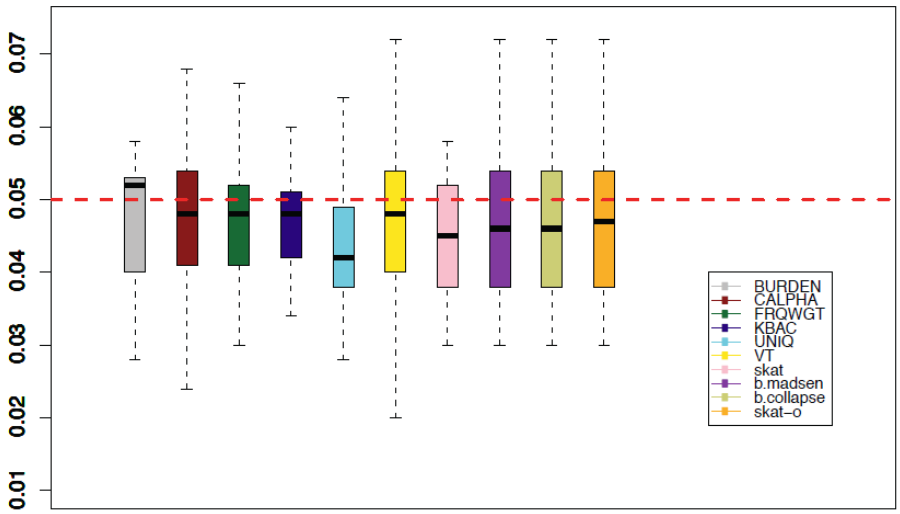


Figure 4.3: Type 1 error of gene-based association methods at alpha=0.05

From **Figure 4.4**, we made several observations. First, at nominal level of significance (α=0.05), many methods have high power (~75%-95%) to detect loci at which deleterious variants (AR1-AR5) explain ~1% of phenotypic variance. KBAC is consistently the most well-powered method to detect deleterious effects at less stringent levels of significance (up to 95% power at α=0.05, under AR4). This high power can be used to detect signals when a small number of hypotheses are being tested (e.g. sequencing across only a few targeted loci), or to confidently exclude rare variant models at the majority of loci tested in a larger-scale scan.

Second, in 3K samples, mean power at an exome-wide significance threshold of α=2.5e-6 (α=0.05, after Bonferroni correction for ~20,000 genes) is very low (5-20%) across all architectures and tests. At a less stringent threshold of α=1e-4, which could be appropriate to nominate loci for further follow-up (under the null, only ~2 such genes are expected exome-wide), mean power of the best performing tests across AR1-AR5 still remains low (10-50%). This is true irrespective of the allele frequency threshold used for inclusion of variants (data not shown).

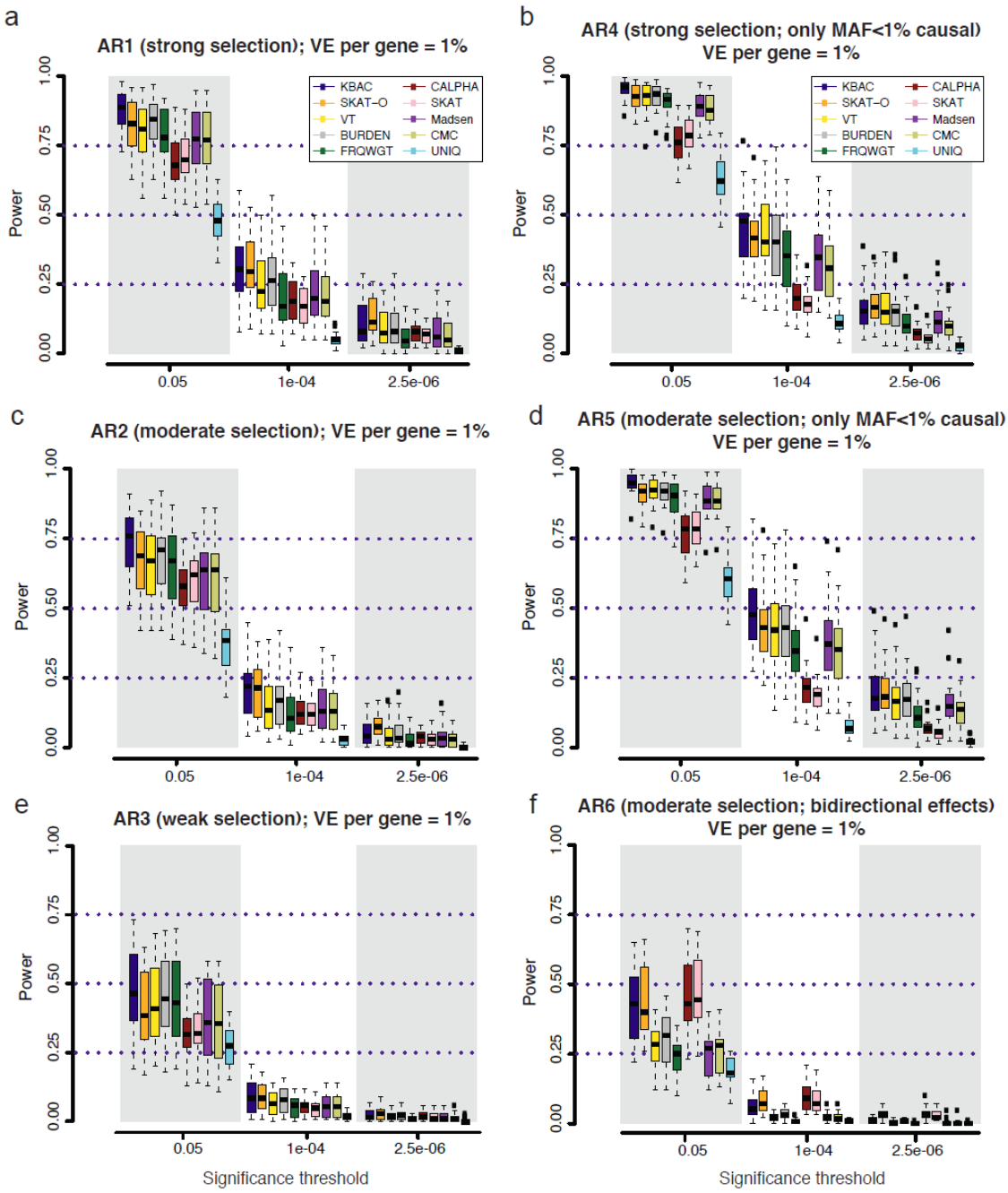


Figure 4.4: Power of different gene-based rare variant association methods at simulated disease loci

At each gene locus, 100 independent simulations of phenotypic effects were generated in a sample size of 3K individuals (1.5K cases / 1.5K controls). Variant effects were drawn from varied models of genetic architecture (**a-f**). At all loci, genetic variants together contribute 1% of the phenotypic variance underlying a trait with common prevalence (8%; modeled as type 2 diabetes). Power is measured as the fraction out of 100 simulations of each gene in which a gene-based test reported a p-value lower than the significance threshold. In (**a-c**), causal variants span the full frequency spectrum (including common alleles), and thus rare alleles account for only a partial fraction of the locus heritability; in (**d-e**), all causal variants are rare (MAF<1%). In (**f**), causal variants have bi-directional effects (some increase risk of disease, while others reduce risk).

Third, some methods appear to be uniformly more powerful than others, independent of locus architecture. Across all architectures in which causal variants have unidirectional (deleterious) effects (**Figure 4.4a-e)**, KBAC and SKAT-O consistently achieve the highest mean power, while UNIQ is least-powered. We do, however, observe differential behavior of these tests depending on the significance threshold; under all architectures, SKAT-O and the variance-component tests (C-ALPHA, SKAT) retain greater power than unidirectional alternatives at stringent thresholds (α<1e-05); at less conservative thresholds (α>1e-03), the opposite is true and KBAC is better-powered than SKAT-O (**Figure 4.4a-f**).

Fourth, the allelic architecture is an important determinant of the power of the best-performing methods. Unsurprisingly, power is uniformly higher when the rare variants included in the association test (e.g. those with MAF<1%) contribute the majority of the locus' effect. This is evidenced by the gain in KBAC power from AR3 (9% at α=1e-4 in 3K samples) to AR2 (22%) to AR1 (32%). Power is higher still under architectures where variants with MAF<1% (i.e. those variants tested) contribute *all* of the locus' effect (AR4 and AR5): KBAC power increases to >45% at α=1e-4 and >90% at α=0.05. Power also depends on the direction of causal effects at a locus: under AR6 (where both risk and protective effects are present), variance-component tests (C-ALPHA, SKAT) and SKAT-O outperform all other methods (as expected[16,20,21]), retaining ~10% power at α=1e-4, while KBAC power is reduced to ~5% (**Figure 4.4f**). As before, SKAT-O is better-powered at stringent significance thresholds; here, C-ALPHA and SKAT are optimal at a nominal threshold.

We next asked how the power of gene-based methods compares to single variant association. In direct contrast to gene-based methods, the power of single variant association tests *decreases* as the contribution of rare variants increase: power at α=5e-08 is ~20%, 10%, and ~7% under AR3, AR2, and AR1, respectively. In all cases, however, the joint application of gene-based and single variant methods yields greater power than single variant association alone (**Figure 4.5a-c**). As expected, the comparative advantage of gene-based tests is most evident under

architectures where there is strong purifying selection against causal alleles (under AR4, for example, the power of single-variant tests at α=5e-8 is <5%, while gene-based tests achieve ~45% power at α=1e-04, and ~20% power even at α=2.5e-06). Under AR3 (where limited purifying selection makes causal alleles more common), gene-based methods have lower absolute power than single variant association, but because different loci are detected by each, power is still maximal when both are applied together.

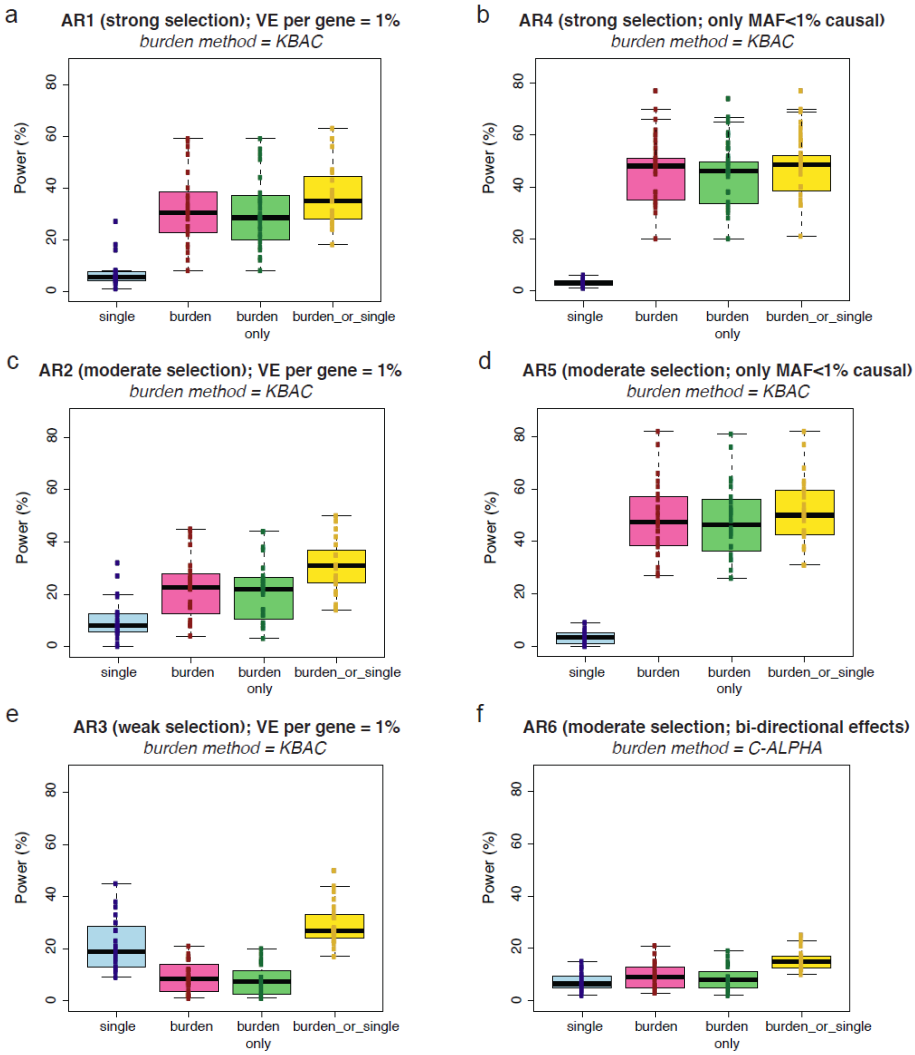


Figure 4.5: Power of best-performing gene-based method as compared to single variant association

Power is measured across 100 simulations of phenotypic effects in N=3K samples (as in Figure 4.4). Under each architecture (**a-f**), power of the best-performing gene-based test is compared to single variant association. The significance threshold used for the gene-based test is 1e-04; the threshold for single variant association (Fisher's exact) is 5e-08. Blue boxplot shows range of power for single variant association across genes simulated; pink shows power of the gene-based test; green shows the fraction of loci detected only by the gene-based test (and not single variant association); yellow shows combined power of both gene-based and single variant association.

In order to characterize the impact of locus effect size on the power of gene-based tests, we simulated loci where the phenotypic variance explained (VE) by genetic variants is 0.5%, 1% (as in **Figures 4.4-5**), and 2% (all under AR2). Power at loci where VE=2% increases to nearly 40%, as compared to ~22% when VE=1% (**Figure 4.6**). When VE=0.5%, power is extremely low (<5% at α=1e-04 in 3K samples), indicating that exome-wide sequencing studies of this size are substantially under-powered to interrogate genes for weak effects. It is worth noting that when VE=0.5%, KBAC outperforms all other methods at α=0.05 by a wide margin, and may thus be the method of choice to nominate loci for putative weak effects (data not shown).

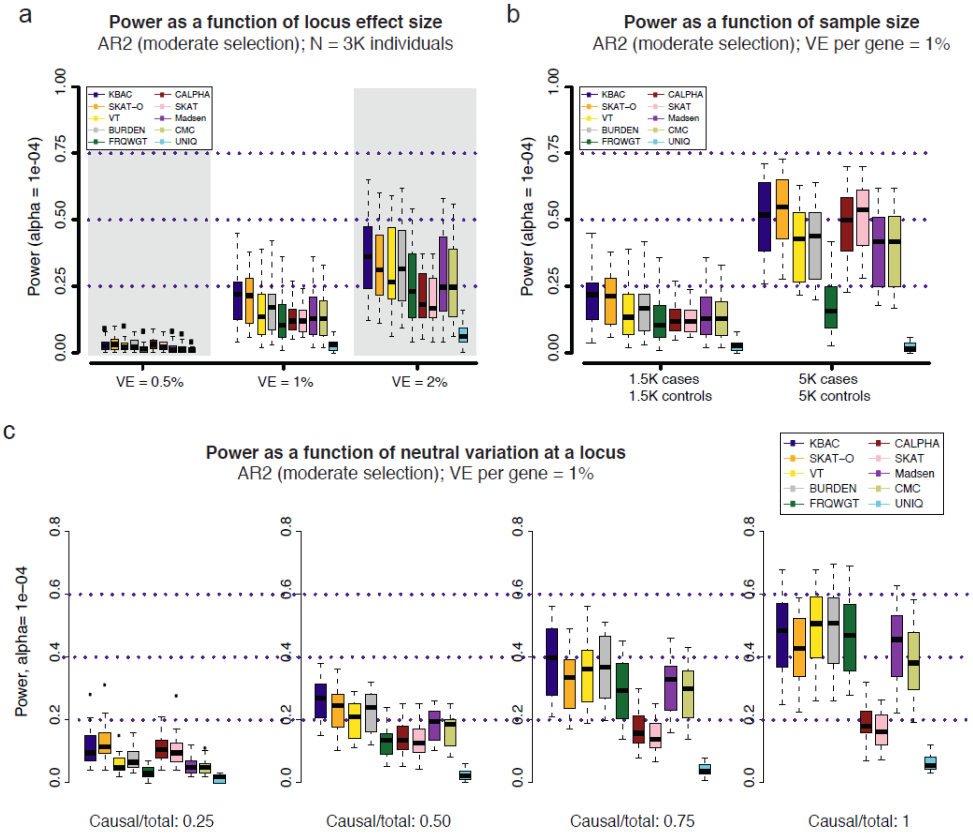


Figure 4.6: Power as a function of locus effect size, sample size, and neutral variation

Power was measured across 100 simulations at each of 24 gene loci (as in Figures 4.4-5). Across all panels above, variant effects were drawn from the model AR2 (assuming moderate selection against causal variants, and thus modest inverse correlation between variant frequency and effect size). In **(a)**, variant effects were sampled at each locus such that the total fraction of phenotypic variance explained by the locus was ~0.5%, 1% (as in Figures 2 and 3) or 2%. In **(b)**, loci were simulated to explain 1% of phenotypic variance in sample sizes of 1.5K cases/1.5K controls and 5K cases/5K controls. In **(a)** and **(b)**, all exonic variants with MAF < 1% were included in the burden test (both causal and non-causal variants, resulting in fewer than 50% of all tested variants being causal). In **(c)**, non-causal (neutral) variants were selectively removed such that the ratio of causal variants to total variants tested ranged from 0.25 to 1 (only causal variants tested).

The relatively modest power of gene-based tests at stringent levels of significance presents challenges to investigators seeking to discover novel disease-associated loci in large-scale studies. Thus, we next investigated the extent to which power could be improved by: a) increasing sample size, b) excluding neutral variation at a locus, or c) selecting a combination of methods best powered to test a particular hypothesis.

We find that some, but not all, methods exhibit substantial gains in power as sample size increases from 3K to 10K individuals (**Figure 4.6**). Median SKAT-O power, for example, increases from ~22% to >55% (at α=1e-04) and remains high (~45%) even at α=2.5e-06. However, the increase in power is not uniform across methods. This occurs because (unlike for single variant tests) the relationship between sample size and power is not straightforward for gene-based tests: as sample size increases, causal alleles are observed more times, but the number of (rare) *non-causal* alleles also grows sharply. As a result, methods that up-weight all rare alleles regardless of their observed effect (e.g., FRQWGT) do not necessarily benefit from larger sample sizes.

The power of gene-based tests is highly sensitive, as has been described[10,11,21], to the fraction of neutral variation at a locus. Our study (a) confirms that power increases as the fraction of neutral variants at a locus decreases, and (b) shows that unidirectional burden tests (and to a lesser extent, SKAT-O) exhibit the sharpest increases in power: in 3K samples, KBAC power increases to >50% (from ~22%) when only disease-causing variants are included (**Figure 4.6c**). These tests, therefore, are preferable for testing targeted hypotheses about a subset of genic variation where rich functional annotation is available. Conversely, variance-component tests as well as SKAT-O are characterized by relative immunity to neutral variation, and are therefore attractive options when jointly testing large numbers of less strictly filtered variants (e.g. in a pathway-based analysis).

Concordance between the results reported by different gene-based association methods

We next investigated the concordance between the results of different gene-based methods to assess the degree of overlap between signals detected by each method. For each pair of

association methods, we computed Pearson's correlation coefficients between their reported p-values on a logarithmic scale (**Figure 4.7a-b**). We find that although tests with similar design characteristics (e.g., SKAT and C-ALPHA) exhibit very high correlation, other methods show varying concordance (**Figure 4.7c**). Some methods are highly correlated, but there is variability in the p-values reported (e.g., KBAC and BURDEN, R^2=0.89). Other methods, such as SKAT and SKAT-O, show asymmetric concordance (R^2=0.78): SKAT-O detects a set of causal loci entirely undetected by SKAT, but is more conservative on the whole, reporting p-values up to an order of magnitude higher than those reported by SKAT at the majority of loci tested. These correlations are also architecture-dependent: under AR2 (where there are only deleterious effects), for example, SKAT-O exhibits highest concordance with KBAC (R^2=0.86), while under AR6 (where bidirectional effects are present), SKAT-O is most concordant with C-ALPHA and SKAT (R^2=0.93). This behavior reflects the 'unified' design of SKAT-O as a combination of a unidirectional burden test and a bidirectional variance-based method.[21]

Finally, some pairs of gene-based methods are much less related (e.g., C-ALPHA and VT, R^2=0.29) or even uncorrelated (e.g., SKAT-O and UNIQ; R^2=0.02). While in this latter case the low correlation is driven by lower mean power of UNIQ relative to SKAT-O, there do exist a subset of true causal loci at which UNIQ reports p<1e-04, but SKAT-O reports p>0.01.

To understand the drivers of such differences and identify scenarios where certain tests may be more powerful than others, we focused on pairwise comparisons between KBAC (one of the highest performing methods at α=1e-04 across AR1-AR5) and the other gene-based methods. For each comparison, we characterized the properties of loci at which KBAC (but not the other method) reports p<0.01, and vice-versa. In the comparison between KBAC and C-ALPHA (**Figure 4.8a**), we find that loci at which only KBAC detects signal are characterized by a higher aggregate skew in case to control counts (often driven by singletons, which do not contribute to the C-ALPHA dispersion statistic). Loci at which only C-ALPHA detects signal, on the other hand, are

characterized by a relatively common single variant of large effect (in the background of many variants with balanced case to control counts).

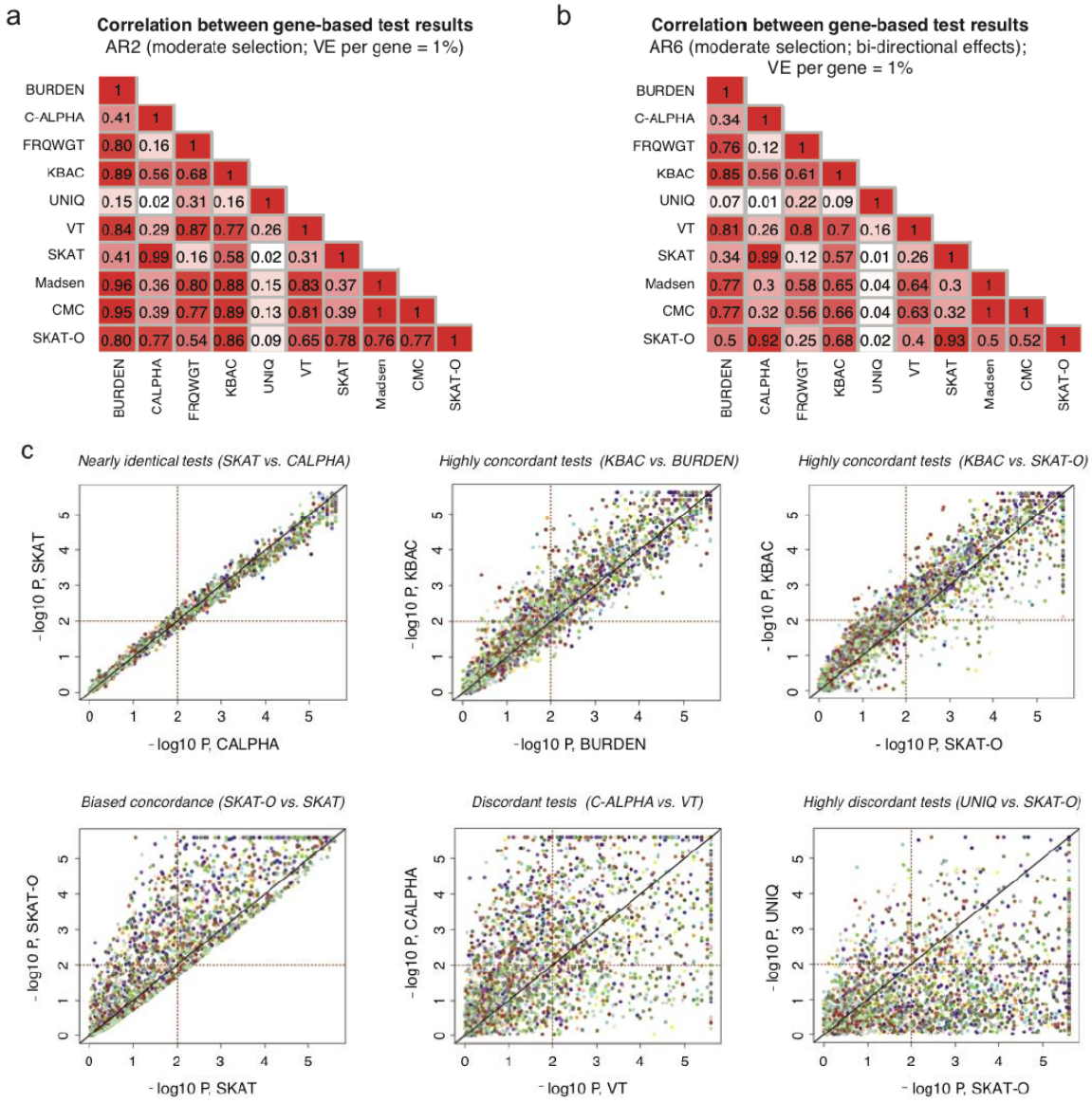


Figure 4.7: Concordance between results of different gene-based methods

Pairwise correlation coefficients (R^2) between the p-values reported by different gene-based association methods under AR2 (moderate selection and unidirectional effects; shown in **a**) and under AR6 (moderate selection and bi-directional effects, shown in **b**). P-values above 0.1 are excluded in computation of the correlation. In c, scatter plots show the results (-log10 of the p-values) reported by a pair of gene-based tests under AR2; p-values lower than 5e-06 are plotted at 5e-06. Each point represents an individual locus at which both gene-based methods were applied (2400 total points); points of the same color represent different simulations at the same human gene locus (e.g. same gene and haplotype structure, but different sampling of variant phenotypic effects). Dotted lines mark p=0.05, such that points above the horizontal line or to the right of the vertical line represent loci at which nominally significant results are reported by the gene-based methods. All data above were generated in 3K samples (1.5K cases and 1.5K controls).

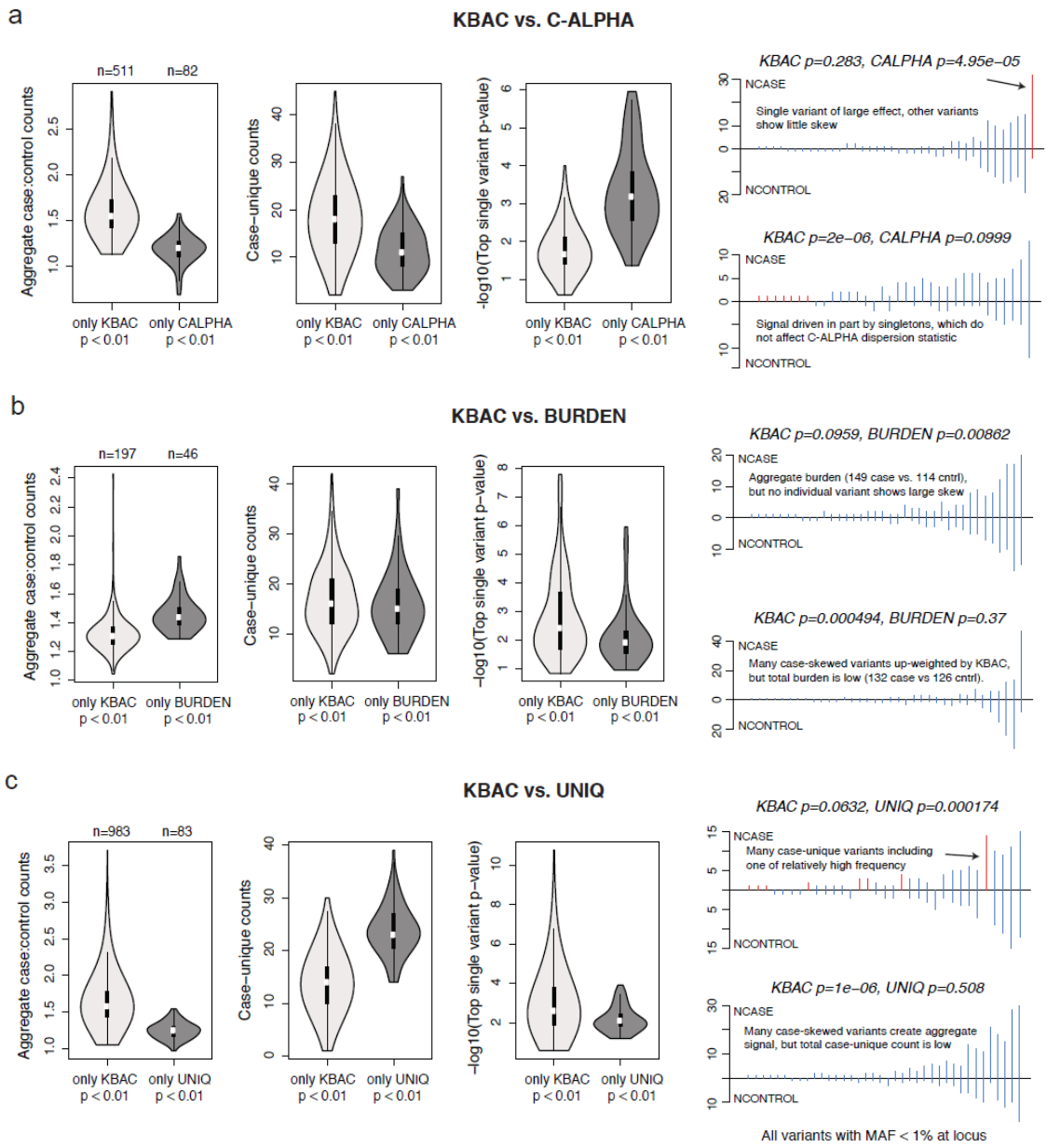


Figure 4.8: Properties of loci at which gene-based methods report discordant results

Characteristics of causal loci at which KBAC (one of the methods with highest mean power at nominal levels of significance) produces discordant results as compared to another gene-based method. KBAC is compared to the **(a)** C-ALPHA, **(b)** BURDEN, and **(c)** UNIQ gene-based methods. In each pairwise comparison, loci are identified at which KBAC (but not the other method) reports a p-value < 0.01, or at which the other method (but not KBAC) reports a p-value < 0.01. For each group of loci, leftmost violin plot shows the distribution of aggregate case to control counts (number of minor alleles observed in cases divided by number of minor alleles observed in controls, for variants with MAF<1%). Middle violin plot shows distribution of case-unique counts (number of observations of alleles that are only present in cases and absent from controls). Rightmost violin plot shows distribution of the top single variant p-value observed for an exonic variant at the locus (log10 scale). Line plots at right show the distribution of variants (MAF < 1%) at representative simulated loci where the methods are discordant. Each line represents a variant; height above line measures the variant's case counts, while height below measures control counts.

For loci where the aggregate case to control count ratio is high, but no individual variant shows any substantial skew, the BURDEN test may be more effective than KBAC (**Figure 4.8b**). This makes sense: KBAC adaptively weights individual variants by their observed case-bias, and if all variants have low weights, the maximum KBAC statistic achievable is low, whereas BURDEN quantifies the significance of the observed signal in *aggregate*. Finally, UNIQ (unsurprisingly) more readily detects loci at which signal is driven by either many rare variants private to cases, or by a single relatively frequent case-unique variant (**Figure 4.8c**). Taken together, these data indicate that although a given method may exhibit high *mean* power across divergent architectures, it can be less powerful than others for testing specific genetic hypotheses.

Summary of findings

This simulation study provides a number of insights informative for the interpretation of ongoing complex disease sequencing studies. Given the low power of single variant association methods to detect rare causal alleles, we confirm that the application of gene-based methods increases power to detect loci at which rare variants drive the causal architecture. In 3K case-control samples, however, we find that the power of gene-based methods to detect loci explaining ~1% of the phenotypic variance underlying a common trait such as type 2 diabetes is limited at stringent levels of significance (~5-20% at α=2.5e-06, and ~10-50% at α=1e-04; power exceeds ~80% only at α=0.05). Even in 10K case-control samples, power of the best-performing methods (SKAT-O, SKAT, KBAC, and C-ALPHA) does not exceed ~60% at α=1e-04; in fact, the increasing number of neutral (non-causal) rare variants in large sample sizes limits the gains in power of many methods (e.g. FRQWGT). Thus, irrespective of the specific locus architecture, we expect that re-sequencing studies will require in excess of 10K samples in order to detect (at very low rates of false discovery) disease loci of modest to large effect size.

Given this low mean power, it is important to identify which particular methods are optimally powered a) to detect causal loci in hypothesis-free settings, across a wide range of architectures, and b) to test specific hypotheses about locus architectures. We find that at stringent significance

thresholds (α<1e-04), SKAT-O is the best-powered method across many architectures, especially when rare variants have bidirectional effects on disease. For investigators looking to discover signals with high specificity across thousands of loci (e.g., in exome-wide scans where dense functional annotation is unavailable), SKAT-O may be the optimal choice.

KBAC, on the other hand, is consistently best-powered to detect rare variants of deleterious effect at less stringent levels of significance (up to 95% power at α=0.05 in 3K samples), and also shows the greatest gain in power when neutral variation is excluded. This attribute may be useful in various scenarios: to test a small number of biological hypotheses (e.g. at only a few loci, especially if functional annotations are available), to prioritize signals for further follow-up from a discovery scan, to confidently exclude genetic models at a majority of re-sequenced loci, or to place bounds (e.g., after an exome-wide sequencing study) on the total number of genes harboring rare variants of a given effect size that are likely to exist. In addition to SKAT-O and KBAC, we find that other methods have individual strengths under particular scenarios (e.g. UNIQ to test whether a gene harbors an excess of penetrant rare variants, or BURDEN to detect a collection of variants each of very weak effect); these methods should be employed to test such specific hypotheses.

In summary, we find that specific gene-based association methods are best deployed in the setting of particular experimental study designs, and to test for particular genetic models of disease. Such an approach will likely enable meaningful interpretation of both positive and negative findings in ongoing sequencing studies, and is bound to remain important even as sample sizes increase and new statistical methods for aggregate testing of rare variants are developed.

References

1. Purcell, S., Cherny, S.S. & Sham, P.C. Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics* **19**, 149-50 (2003).
2. Kiezun, A. *et al.* Exome sequencing and the genetic basis of complex traits. *Nature Genetics* **44**, 623-30 (2012).
3. Asimit, J. & Zeggini, E. Rare variant association analysis methods for complex traits. *Annual Review of Genetics* **44**, 293-308 (2010).
4. Stitziel, N.O., Kiezun, A. & Sunyaev, S. Computational and statistical approaches to analyzing variants identified by exome sequencing. *Genome Biology* **12**, 227 (2011).

5. Rivas, M. a *et al.* Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nature Genetics* **43**, 1066-73 (2011).
6. Cohen, J.C., Boerwinkle, E., Mosley, T.H. & Hobbs, H.H. Sequence Variations in PCSK9, Low LDL, and Protection against Coronary Heart Disease. *New England Journal of Medicine* 1264-1272 (2006).
7. Johansen, C.T. *et al.* Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. *Nature Genetics* **42**, 684-7 (2010).
8. Bonnefond, A. *et al.* Rare MTNR1B variants impairing melatonin receptor 1B function contribute to type 2 diabetes. *Nature Genetics* **44**, 297-301 (2012).
9. Bansal, V., Libiger, O., Torkamani, A.L.I. & Schork, N.J. An application and empirical comparison of statistical analysis methods for associating rare variants to a complex phenotype. *Pac Symp Biocomput* 76-87 (2011).
10. Ladouceur, M., Dastani, Z., Aulchenko, Y.S., Greenwood, C.M.T. & Richards, J.B. The empirical power of rare variant association methods: results from Sanger sequencing in 1,998 individuals. *PLoS Genetics* **8**, e1002496 (2012).
11. Basu, S. & Pan, W. Comparison of Statistical Tests for Disease Association with Rare Variants. *Genetic Epidemiology* **35**, 606-619 (2011).
12. Su, Z., Marchini, J. & Donnelly, P. HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics* **27**, 2304-5 (2011).
13. The 1000 Genomes Project Consortium A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-73 (2010).
14. Nelson, M. *et al.* An Abundance of Rare Functional Variants in 202 Drug Target Genes Sequenced in 14,002 People. *Science* **337**, 100-104 (2012).
15. PLINK/SEQ: A library for the analysis of genetic variation data. at <http://atgu.mgh.harvard.edu/plinkseq/>
16. Neale, B.M. *et al.* Testing for an unusual distribution of rare variants. *PLoS Genetics* **7**, e1001322 (2011).
17. Li, B. & Leal, S.M. Methods for Detecting Associations with Rare Variants for Common Diseases : Application to Analysis of Sequence Data. *The American Journal of Human Genetics* 311-321 (2008).doi:10.1016/j.ajhg.2008.06.024.
18. Liu, D.J. & Leal, S.M. A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. *PLoS Genetics* **6**, e1001156 (2010).
19. Madsen, B.E. & Browning, S.R. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genetics* **5**, e1000384 (2009).
20. Wu, S. *et al.* Rare Variant Association Testing for Sequencing Data Using the Sequence Kernel Association Test (SKAT). *American Journal of Human Genetics* **89**, 82-93 (2011).
21. Lee, S. *et al.* Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *American Journal of Human Genetics* **91**, 224-37 (2012).
22. Price, A.L. *et al.* Pooled association tests for rare variants in exon-resequencing studies. *American Journal of Human Genetics* **86**, 832-8 (2010).
23. EPACTS: Efficient and Parallelizable Association Container Toolbox. at <http://genome.sph.umich.edu/wiki/EPACTS>
24. So, H.-C., Gui, A.H.S., Cherny, S.S. & Sham, P.C. Evaluating the heritability explained by known susceptibility variants: a survey of ten complex diseases. *Genetic Epidemiology* **35**, 310-7 (2011).
25. Falconer, D.S. The inheritance of liability to diseases with variable age of onset, with particular reference to diabetes mellitus. *Annals of Human Genetics* **31**, 1-20 (1967).

Chapter 5

Locus architecture at the T2D-associated chr9p21 non-coding region

The first four chapters of this thesis leveraged simulation-based studies to model the global and locus genetic architecture of complex human diseases. Simulations are only useful in bounding genetic architecture, however, when they can be compared directly to *observed* genetic data; deep characterization of empirical data will always remain the most critical part of the approach proposed in Chapters 1-3. Moreover (as alluded to at the end of Chapter 3), constraints on global genetic architecture are not actually sufficient to exclude *locus-specific allelic architectures* at individual disease loci. Even if the global architecture of type 2 diabetes (T2D) was most consistent with a common polygenic model, for example, models implicating rare causal variation could still exist at a subset of disease loci. Such models have been especially difficult to interrogate in empirical data, however, because rare genetic variation had not (until recently) been discovered, genotyped, or characterized in large sample sizes.

We were fortunate to have the opportunity to analyze data from a number of large-scale genetic studies that were undertaken for type 2 diabetes in the past few years (see Background). Amongst these was a low-pass whole-genome sequencing study of ~2800 unrelated European cases (with type 2 diabetes) and controls. These data provided unprecedented (and near-complete) testing of both common and low-frequency genetic variation (in both protein-coding and non-protein-coding regions of the genome) for association to risk of T2D. Although power to discover novel loci is relatively low in this sample size (as seen in **Figure 3.8**), these data provided a unique opportunity to systematically characterize the allelic architecture at previously known T2D GWAS loci (most of which fall in non-protein-coding regions of the genome).

In this chapter, we focus on the non-coding chr9p21 locus, which harbors one of the strongest known common variant GWAS signals for T2D. **Appendix A2** contains a manuscript we published in 2011, which compared genotyping and imputation strategies for fine-mapping at this

locus. In this chapter, we extend this work and describe haplotype-based methods for fine-mapping and genetic hypothesis testing using complete sequence data. We use these methods to nominate a set of candidate causal variants and exclude the possibility of synthetic associations at chr9p21. In Chapter 6, we perform similar analyses to characterize architecture at 10 other T2D GWAS loci.

Background: fine-mapping of GWAS loci

The causal variant(s) underlying signals discovered in genome-wide association studies (GWAS) for complex traits are, in most cases, unknown. The allelic architecture at these loci – the number, frequencies, and effect sizes of causal mutations – is not yet understood, in part because GWAS directly tested only a subset of common marker polymorphisms.[1] Identification of the causal class is a key step towards elucidating both the inheritance patterns and the biological mechanisms of common human diseases.

A range of genetic models have been hypothesized to explain observed GWAS signals; these include common (MAF>5%) causal variants of weak effect, individual low frequency (1%<MAF<5%) variants of large effect, a burden of rare (MAF<1%) variants with aggregate effect on disease, or some combination of these. The latter two genetic models, in which low frequency and rare causal variants are assumed to segregate on a subset of haplotypes carrying a GWAS disease-associated marker allele, have been described as models which could produce 'synthetic' common variant associations[2–6]. Finally, it has also been hypothesized that common variants alone may underestimate the total contributions of GWAS loci to disease heritability, and that additional causal variants (independent of the original GWAS association signals) may exist at these loci.[7,8]

In order to systematically test these hypotheses about allelic architecture at any GWAS locus, it is critical to (a) identify and genotype all segregating genetic variation (both common and rare) across the region in a *phenotyped* population sample, (b) test each variant for association to disease, both individually and in combination, (c) characterize the haplotypes on which GWAS tag SNPs originally produced an association signal, identifying all the genetic variants that tag or

partially tag these haplotypes, and (d) enumerate the plausible genetic models that would be consistent with these data. The genetic models that remain can then be further evaluated for their statistical likelihood of explaining the observed data (e.g., in larger sample sizes), and the candidate causal variants implicated in each model can then be tested experimentally (e.g., in biological systems).

Background: the chr9p21 T2D GWAS locus and previous fine-mapping efforts

Association of common variants at chr9p21 to risk of T2D was first reported in multiple GWAS conducted in 2007.[9–11] Since then, other variants across the broader locus have been associated with a host of human phenotypes, including myocardial infarction[12], aneurysm, vertical cup disc ratio, glaucoma, and multiple cancers (leukemia[13], breast cancer, melanoma, basal cell carcinoma, glioma). The variants associated with T2D are independent (not in linkage disequilibrium with) these other variants, and are in fact confined to an (unusually small) ~10kb region that lies between two strong hotspots of recombination (**Figure 5.1**). This region contains no protein-coding genes, and is over 100kb away from the nearest protein-coding genes (the tumor suppressors CDKN2A/B), suggesting that causal variants within this small region may alter the regulation of gene expression.[14]

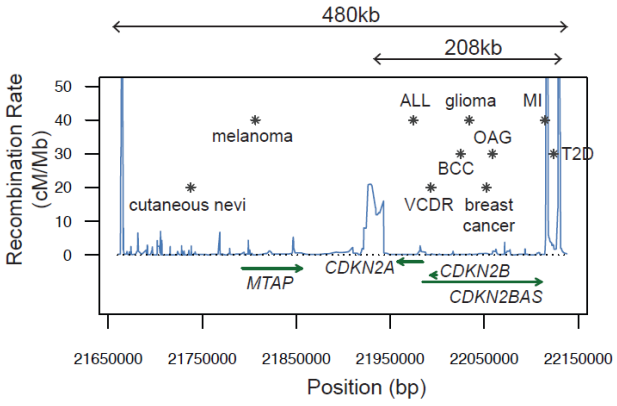


Figure 5.1: Common variant GWAS signals across the chr9p21 locus (Source: Jessica Alston)

The T2D signal is confined to a small ~10kb region between two strong hotspots of recombination.

Further adding to the intrigue of the chr9p21 locus is the fact that there is strong evidence for *multiple* independent association signals within this ~10kb linkage block. Two independent signals have been reported to create a three-tiered haplotype association; haplotypes can be classified as risk, neutral, or protective (reflecting the risk they confer for T2D) based on genotype at two common marker SNPs.[15,16] The causal variants driving these association signals, however, remain unknown.

Previous fine-mapping studies at the 9p21 locus were limited by incomplete catalogs of genetic variation, limited imputation accuracy (especially for rare variants), and genotyping assay failure. To date, two main fine-mapping efforts have been undertaken at this locus. The first (see **Appendix A2** of this thesis)[16] leveraged high-coverage targeted sequencing in 47 controls to assemble a catalog of intermediate frequency and common variants, followed by imputation into ~2K individuals phenotyped for T2D. This study identified ten common SNPs (indistinguishable in this sample size) tagging the protective haplotype, but none tagging the risk haplotype. This study also demonstrated that imputation quality increases dramatically when samples are densely genotyped (e.g., beyond GWAS array density). The second[17], more recent, fine-mapping effort at 9p21 attempted direct genotyping of a panel of SNPs (including those our group detected in the first effort, as well as those discovered by the 1000 Genomes Project) across the T2D region in ~5K phenotyped individuals. This study – again, limited to common and only some low frequency variants – identified a set of common SNPs likely to explain the GWAS signal. Genetic hypotheses about low frequency and rare variants were largely untested by these studies.

Results of analyses conducted in large panel of sequenced case-control samples

Here, we analyzed data from a low-pass whole-genome sequencing study of ~2,800 individuals of mixed European ancestry sampled from extremes of the phenotypic distribution (GoT2D Consortium; 2,657 samples after quality control). We selected a 5Mb region at chr9:20,000,000-25,000,000 (hg19) and identified 55,359 total segregating variants (both SNPs and indels) across this region; 171 of these variants lie within the T2D GWAS locus (chr9:22,125,000-22,142,000, a ~17kb region including the flanking recombination hotspots). We confirmed that sequencing coverage was relatively uniform across this region, with >90% of samples having a median read depth of >=4 per base (**Figure 5.2**). As expected, most segregating variants are rare: 52 of the 171 variants were seen only once across all the sequenced samples, and an additional 34 variants had MAF< 0.1%. A limitation of low-pass sequencing is that we had only modest sensitivity

to detect such extremely rare variants (~30% sensitivity for singletons, and ~60% for non-singleton variants with MAF<0.1%).

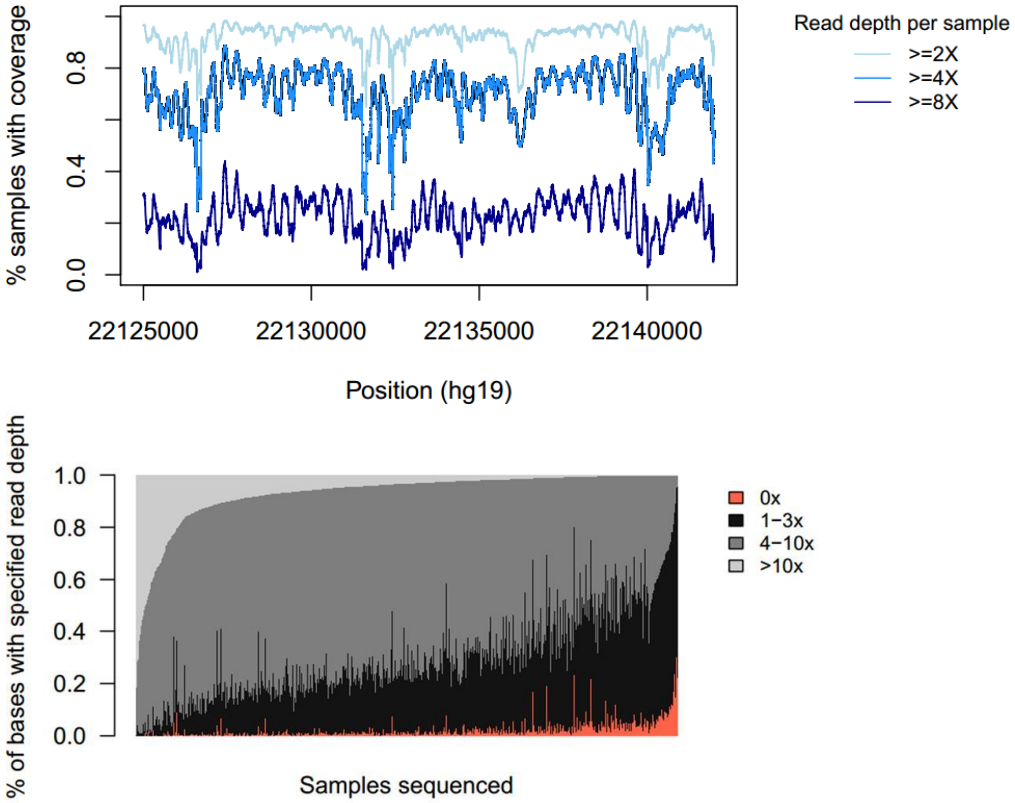


Figure 5.2: Sequencing coverage across the chr9p21 locus

The chr9p21 locus was sequenced as part of a low pass whole genome sequencing study. Sequencing coverage across this locus **(a)** was relatively uniform, with ~75% of samples consistently having at least 4x read coverage at each position. The second panel **(b)** shows each sample across the x-axis; the y-axis shows the percent of bases across the chr9p21 at which the individual was sequenced at 0x, 1-3x, 4-10x, or >10x depth.

Nonetheless, this set of variants represents the largest catalog of segregating genetic variation across the non-coding 9p21 locus to date (**Figure 5.3a**). In particular, we estimate that we have identified >90% of all genetic variants with MAF between 0.1-1%, and all variants with MAF>1%. This is in stark contrast to the HapMap Phase 2 and 3 catalog, which includes only ~65% of common (MAF>5%) variants, and no representation of lower frequency variants at the 9p21 locus. The re-sequencing experiment conducted by Shea et al identified ~90% of common variants, and ~60% of MAF 1-5% variants, but <20% of variants in the MAF 0.1-1% class. Even the catalog of

variants reported by the 1000G Project's Phase 1 sequencing of Europeans, while nearly complete for variants with MAF>1%, includes less than half of all variants in the MAF 0.1-1% class.

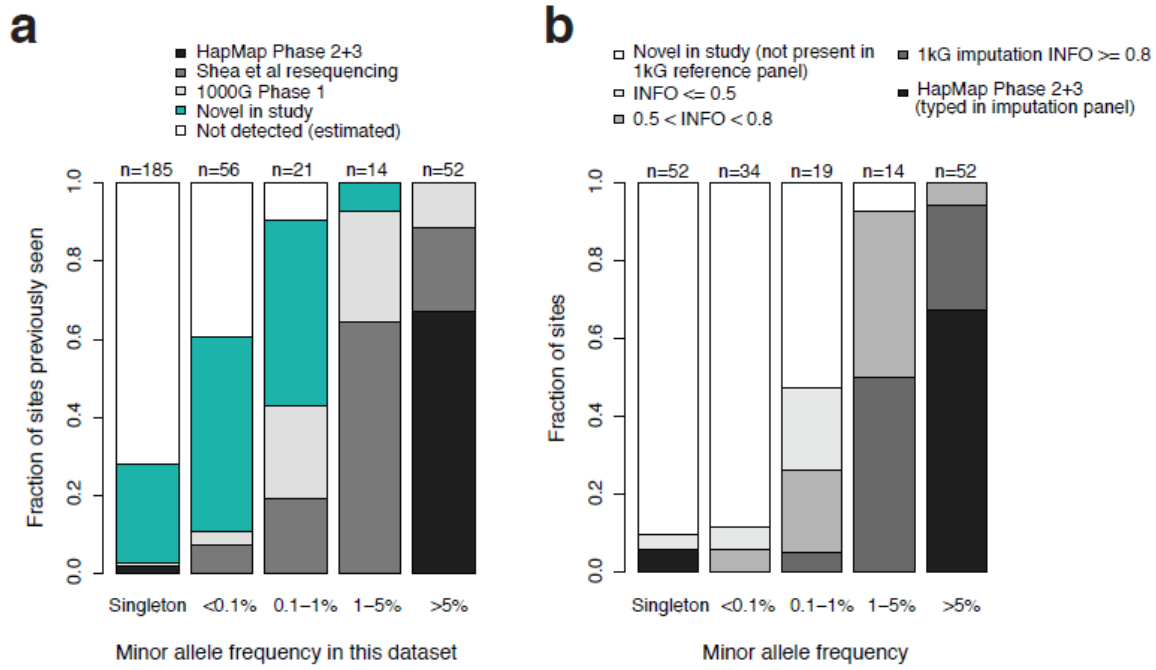


Figure 5.3: Variant catalog at the 9p21 T2D locus resulting from low pass sequencing of 2,657 European individuals, compared to previous catalogs of sequence variation.

(a) Site discovery in this study as compared to HapMap Phase 2 and 3, prior re-sequencing by Shea et al, and the 1000 Genomes Project Phase 1. **(b)** Genotype capture in this study as compared to imputation of variants from the 1000G data. The total number of variants per frequency bin is different from panel (a) because here only variants detected in GoT2D sequencing are included, e.g. white portions of bars in panel (a) are excluded in panel (b). Imputation was performed by down-sampling sequenced GoT2D data to only sites present in HapMap Phase 2 or 3 (black), pre-phasing this panel using SHAPEIT, and finally imputing data from the phased 1000 Genomes project reference panel into the disease cohort using IMPUTE2. INFO score represents the metric reported by IMPUTE2.

We next asked what fraction of variation present in prior catalogs (e.g., from the 1000 Genomes Project Phase 1 data) could have been successfully imputed into our disease samples, had we not undertaken complete sequencing of these samples. In order for a variant catalog to be truly 'complete' with respect to a phenotype of interest, each variant identified must not only be discovered but also be *genotyped* accurately in a panel of phenotyped individuals such that its association to the trait can be tested. This analysis revealed that public variant catalogs are in fact further incomplete for rare and intermediate frequency variants (**Figure 5.3b**). When attempting imputation of 1000G Project variants into our panel of 2,657 sequenced samples (down-sampled to only those sites in HapMap Phase 2+3), we found that only about half of all variants with MAF 1-5%

and about a tenth of variants with MAF 0.1-1% could be imputed with an estimated r^2 >= 0.8. Combined with incompleteness of the 1000 Genomes reference panel, this resulted in a genotype capture rate of ~50% among variants with MAF 1-5% and <5% among variants with 0.1-1%. This is in contrast to the catalog resulting from direct sequencing of phenotyped samples, which enabled genotype capture of all variants with MAF 1-5% and ~90% of variants with MAF 0.1-1%. Thus, this genetic dataset represents not only the largest but also by far the most *complete* catalog of variation at the chr9p21 T2D locus, and enables (for the first time) nearly comprehensive testing of all variants down to 0.1% frequency for association to risk of T2D.

We first tested all individual variants (SNPs and indels, both common and rare) across a 5Mb region for association to T2D. As expected, the top association signals fall within the previously known ~10kb block of linkage disequilibrium (LD) in which previously reported GWAS marker SNPs lie (**Figure 5.4a-b**). In this region, the most associated variants are common; the top two signals are two closely linked (r^2 = 0.96) SNPs, rs7018475 (OR = 1.28, MAF = 28%, p=4e-05) and rs12555274 (OR = 1.28, MAF = 28%, p=5e-05). Interestingly, neither of these variants was tested in initial GWAS for T2D; rs7018475 is in the HapMap catalog, but was not on first-generation GWAS arrays, and rs12555274 is not present in the HapMap variant catalog at all.

We next asked whether signals independent of these top signals could be detected. After conditioning on rs7018475, we find that a group of common (MAF~12%) SNPs in modest LD with each other (but low LD with rs7018475) show residual signal in the *protective* direction: the lead SNP after conditioning is rs1333051, which shows OR = 0.76 and p=2e-03 (**Figure 5.4c**). These data suggested the presence of multiple independent signals, as was expected based on previous studies.[15,16] Indeed, we confirmed that the previously described three-tiered haplotypic association – using haplotypes defined by the common markers rs10757282 and rs10811661 – is observed in this dataset (omnibus p=1.3e-04; **Table 5.1**).

In order to better understand the relationship between the previously reported GWAS SNPs and the association signals observed in this dataset, we next sought to characterize the disease-

associated *haplotypes* observed at the 9p21 locus. To do this, we phased the genotype data (e.g. inferred haplotypes) across all individuals in the sequenced sample using the software BEAGLE. Focusing on only the 24 common (MAF>5%) variants in the ~10kb region (excluding recombination hotspots), we identified 28 unique haplotypes present at frequency >=0.1% (8 of these were observed >100 times; these are shown in **Figure 5.5**). We then tested each haplotype for association to T2D; groups of haplotypes showing association in the risk and protective direction were immediately evident.

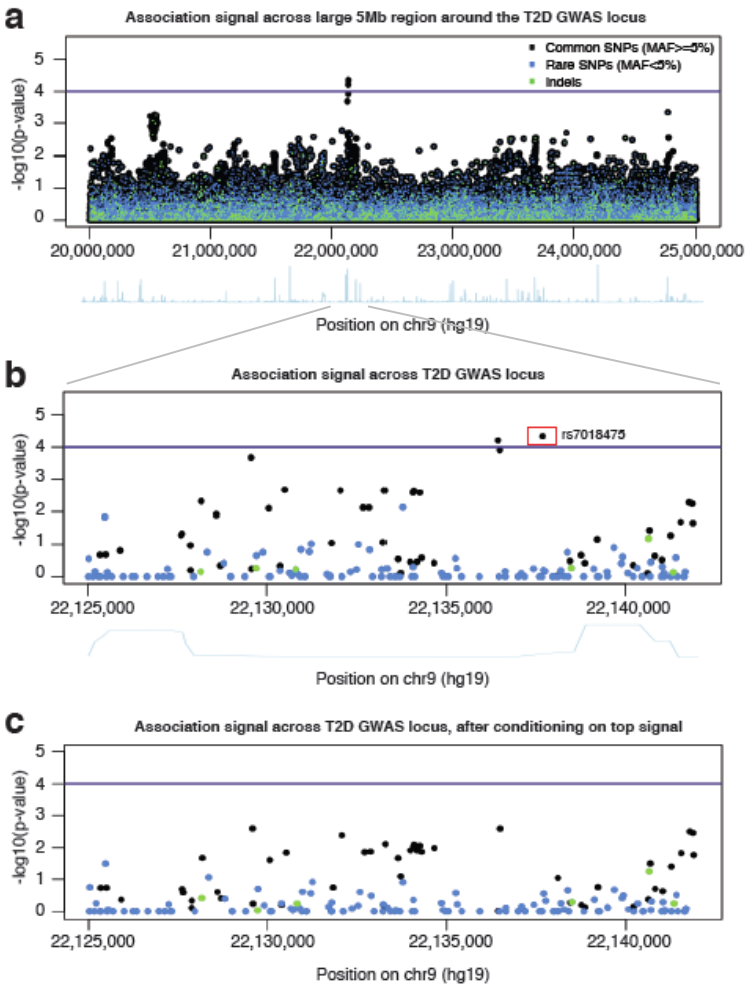


Figure 5.4: Association results for T2D across all variants identified in the 9p21 region

(a) Association results across a 5Mb region (chr9:20,000,000-25,000,000; hg19). **(b)** Association results across the T2D GWAS locus (chr9:22,125,000-22,140,000; hg19; includes recombination hotspots). **(c)** Association results, after conditioning on the top signal in this dataset (rs7018475). Manhattan plot shows -$\log_{10}$ (p-value); light blue line plot below indicates recombination rate across the region. In all plots, black points represent common SNPs (MAF>=5%); blue points represent rare or low frequency SNPs (MAF<5%); green points represent small insertions or deletions (of any frequency).

Table 5.1: Evidence for previously described haplotype association in GO-T2D study of n=2,657 samples

Haplotypes defined by ***rs10757282*** *and* ***rs10811661***

	Frequency			
Haplotype	Cases	Controls	OR	P-value
Overall evidence	-	-	-	**1.3E-04**
TT (*neutral*)	0.54	0.55	0.95	0.28
CT (*risk*)	0.33	0.28	1.23	3.2E-04
CC (*protective*)	0.13	0.16	0.79	1.7E-03

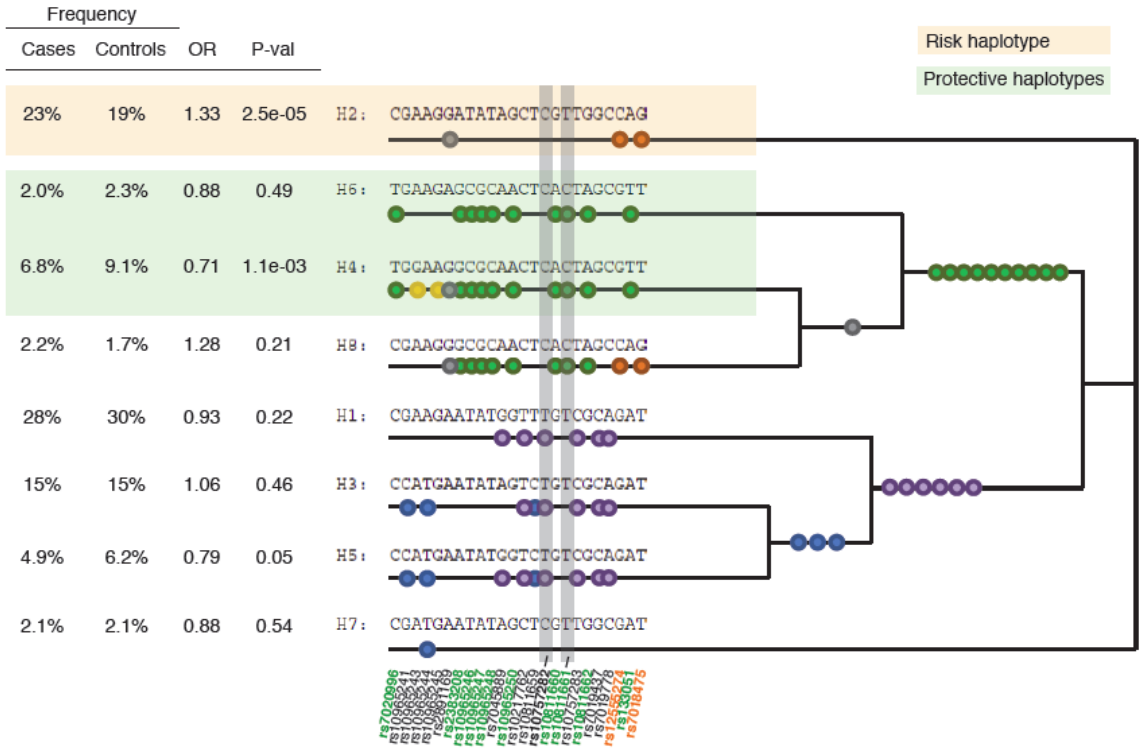


Figure 5.5: Haplotype structure observed in sequenced cases and controls at the 9p21 locus

Haplotypes above were constructed using 24 common (MAF>5%) SNPs located in between the recombination hotspots at the T2D GWAS locus at chrp21 (these are labeled at bottom). The above eight haplotypes were the most common haplotypes observed in the GoT2D dataset; these were each observed >100 times across 2,657 samples. H7 and H8 appear to be recombinant haplotypes.

Each haplotype's frequency among cases and controls as well as the results of association testing between the haplotype's dosage and risk of T2D are shown at left. The two SNPs with grey shading above them are rs10757282 and rs10811661, the SNPs based on which a three-tiered haplotype association has been previously reported (as in **Table 5.1**). SNPs tagging the risk haplotype are labeled in orange, while SNPs tagging the protective haplotypes are in green.

Furthermore, the relationship between different groups of variants was clarified by the haplotype structure. The two top signals in our dataset – rs7018475 and rs12555274 – are the only two SNPs that uniquely define the 'risk' haplotypes (orange in **Figure 5.5**). The group of SNPs

showing signal after conditioning on the risk SNPs (**Figure 5.4c**) are SNPs that tag, to varying degrees, the group of haplotypes that show association in the protective direction (green and yellow in **Figure 5.5**). These SNPs include the previously described GWAS tag SNP rs10811661 as well as the top conditional signal in our dataset (rs1333051). Interestingly, the GWAS marker SNP rs10757282 (previously described as the top independent signal after conditioning on rs10811661) is actually one of many common variants tagging *neutral* haplotypes (which show no significant association to T2D); the two alleles at rs10757282 separate the neutral and risk haplotypes after conditioning on the protective marker rs10811661 (grey shaded rectangles in **Figure 5.5**). This highlights the challenge of interpreting conditional association signals without appropriate haplotype context; a statistically associated signal does not necessarily represent a candidate causal variant.

Given these data, we reasoned that the most parsimonious genetic model for the locus would implicate only two causal variants, one on the risk haplotype and one on the protective haplotype. We thus asked: which variants could alone explain the observed risk and protective haplotype associations? To assemble this set of variants, we added, one by one, the dosage of each variant (common and rare) across the entire 5-Mb region at 9p21 to a logistic regression model of phenotype vs. haplotype dosage. We performed this procedure separately for the risk and protective haplotypes, and recorded the variants which, when added to the model, resulted in no significant effect remaining at the haplotype dosage (e.g. these variants could individually explain the haplotype's association). We find that the only variants that can statistically explain the risk and protective haplotype associations are common SNPs, and they all lie within the small T2D GWAS region (**Table 5.2**). Unsurprisingly, these variants (2 on the risk haplotype, 13 on the protective haplotype) are the same as those visually seen as tagging the haplotypes in **Figure 5.5**.

This analysis did not reveal any rare or intermediate frequency variants of sufficient effect size to *individually* produce 'synthetic' association of the common risk or protective haplotypes; this is perhaps not surprising given the large effect sizes that would be required for an individual rare causal variant to drive association at a common marker. The intermediate frequency variant with the

strongest observed association signal across the 9p21 T2D GWAS region is rs76011118 (chr9:22133772; MAF 3.5%; OR=1.50, p=0.008). This variant occurs on the background of the risk haplotype and thus could potentially be a candidate causal variant, but the haplotype retains significant signal (OR=1.28, p=2.0e-05) even after conditioning on the dosage of rs76011118. Thus, the observed genetic data exclude the possibility of synthetic associations produced by *single* rare or intermediate frequency variants.

Table 5.2a: Single variants sufficient to explain the common risk haplotype signal

All single variants across the 5Mb were tested in a joint regression model with dosage of the common risk haplotype (H2); only the above two common SNPs were sufficient to reduce the association signal at the risk haplotype such that the haplotype *p*>0.05.

Variant	rsID	MAC	MAF	Risk haplotype association p-value after inclusion of single variant
9:22136440	rs12555274	1477	27.8%	0.13
9:22137685	rs7018475	1496	28.2%	0.16

Table 5.2b: Single variants sufficient to explain the common protective haplotype signal

All single variants across the 5Mb were tested in a joint regression model with dosage of the common protective haplotype (H4+H6); only the above 13 common SNPs were sufficient to reduce the association signal at the protective haplotype such that the haplotype *p*>0.05.

Variant	rsID	MAC	MAF	Protective haplotype association p-value after inclusion of single variant
9:22128180	rs12379111	594	11.2%	0.05
9:22129579	rs7020996	611	11.5%	0.59
9:22130065	rs10965243	530	10.0%	0.06
9:22130515	rs10965245	480	9.0%	0.20
9:22132076	rs2383208	838	15.8%	0.16
9:22132698	rs10965246	837	15.8%	0.06
9:22132729	rs10965247	837	15.8%	0.06
9:22132878	rs10965248	841	15.8%	0.06
9:22133284	rs10965250	789	14.8%	0.18
9:22134068	rs10811660	790	14.9%	0.16
9:22134094	rs10811661	791	14.9%	0.18
9:22134253	rs10811662	790	14.9%	0.16
9:22136489	rs1333051	640	12.0%	0.57

It has also been hypothesized that a *collection* of low frequency (MAF<5%) causal variants could, in aggregate, produce association at a common marker if these variants were localized, by chance, to the same common haplotype. More precisely, it is possible that a group of low frequency variants *sharing ancestry* with a disease-associated common marker allele could be driving the marker's association signal. To test this hypothesis, we first used the phased genotype data to place all low

frequency variants, excluding singletons, within the 9p21 T2D locus onto the background of the common haplotypes identified in **Figure 5.5**. Given uncertainty in rare variant phase information, we checked the estimated phase in several ways to ensure that low frequency alleles were consistently observed on the same common haplotype background; in some cases, we manually corrected the BEAGLE-estimated phase. Across the ~17kb 9p21 locus (where a total of 171 variants were seen), we identified 20 low frequency variants (excluding singletons) segregating on the risk haplotype, and 19 variants on the protective haplotype (**Table 5.3**).

Table 5.3: Variants sharing ancestry with common risk and protective haplotypes

Listed above are all rare and low frequency (MAF>5%, seen >1x) variants identified either as a) occurring on the background of the common risk / protective haplotypes (based on phasing of the sequenced data) within the T2D GWAS locus, or b) as sharing long-range ancestry with the risk or protective alleles (using ancestral recombination graphs). Variants sharing long-range ancestry (e.g., falling outside the region flanked by recombination hotspots) are shaded in grey above. Variants are sorted by odds ratio.

Variants sharing ancestry with risk haplotype

					Single variant association statistics	
Variant	MAF	MAC	Case_Obs	Control_Obs	OR	P-value
9:22103887	0.09%	5	5	0	-	-
9:22132115	0.04%	2	2	0	-	-
9:22139928	0.09%	5	4	1	3.75	0.238
9:22120861	0.11%	6	4	2	2.13	0.386
9:22135292	0.06%	3	2	1	2.11	0.543
9:22137433	0.06%	3	2	1	2.04	0.562
9:22133773	3.50%	186	111	75	1.51	0.007
9:22116261	1.00%	53	31	22	1.38	0.259
9:22110965	0.09%	5	3	2	1.37	0.685
9:22129902	1.71%	91	53	38	1.35	0.172
9:22117513	0.21%	11	6	5	1.35	0.623
9:22125548	0.17%	9	5	4	1.27	0.722
9:22112551	0.43%	23	12	11	1.05	0.901
9:22138283	0.04%	2	1	1	1.04	0.978
9:22139497	0.15%	8	4	4	0.97	0.967
9:22131774	0.04%	2	1	1	0.88	0.928
9:22137710	0.79%	42	19	23	0.86	0.618
9:22132897	3.63%	193	86	107	0.80	0.144
9:22130894	0.06%	3	1	2	0.80	0.855
9:22125032	0.96%	51	21	30	0.73	0.271
9:22123005	0.88%	47	20	27	0.73	0.274
9:22125213	0.13%	7	3	4	0.73	0.680
9:22138195	0.09%	5	2	3	0.65	0.636
9:22134071	0.11%	6	2	4	0.55	0.487
9:22130827:indel	0.06%	3	1	2	0.51	0.587
9:22131084	0.06%	3	0	3	-	-
9:22116362	0.04%	2	0	2	-	-

Variants sharing ancestry with protective haplotype

					Single variant association statistics	
Variant	MAF	MAC	Case_Obs	Control_Obs	OR	P-value
9:22140228	0.04%	2	0	2	-	-
9:22131252	0.13%	7	1	6	0.16	0.094
9:22138382	0.15%	8	2	6	0.33	0.173
9:22128810	0.11%	6	2	4	0.46	0.375
9:22123737	0.53%	28	10	18	0.49	0.077
9:22125476	1.17%	62	20	42	0.51	0.014
9:22130827:indel	0.06%	3	1	2	0.51	0.587
9:22134071	0.11%	6	2	4	0.55	0.487
9:22139684	1.37%	73	29	44	0.69	0.125
9:22141685	0.09%	5	2	3	0.70	0.702
9:22130974	2.56%	136	59	77	0.77	0.141
9:22132576	3.73%	198	87	111	0.78	0.097
9:22131202	2.54%	135	59	76	0.78	0.165
9:22132897	3.63%	193	86	107	0.80	0.144
9:22140712	1.62%	86	38	48	0.81	0.349
9:22132356	0.28%	15	7	8	0.82	0.697
9:22131787	0.98%	52	27	25	1.01	0.962
9:22134860	1.07%	57	29	28	1.08	0.764
9:22137765	1.05%	56	29	27	1.12	0.666
9:22128352	2.58%	137	77	60	1.28	0.172

While the strength of the recombination hotspots flanking the 9p21 T2D locus makes it unlikely that long-range haplotypes persist beyond these boundaries, it is possible that some individuals carry longer-range haplotypes. In fact, since the hotspots were inferred based only on common variant genetic data, low frequency alleles *outside* the T2D locus may still share (more

recent) ancestry with the common marker alleles tagging the risk and protective haplotypes. Moreover, it has been hypothesized[2] that recombination may not completely eliminate synthetic associations. To address these issues, we constructed ancestral recombination graphs (ARGs[18]) across the full 5-Mb region centered at the T2D locus (**Figure 5.6a-b**). This analysis enabled us to quantify, for each chromosome in each individual, the length of the haplotype that shared ancestry with a given common marker allele (e.g. the haplotype region estimated to have undergone no recombination since the occurrence of the common allele). The ARGs revealed that while the majority of chromosomes likely underwent recombination very close to the recombination hotspots, a subset of individuals carry haplotypes extending up to ~30kb away. We thus identified an additional 7 variants sharing ancestry with the risk haplotype and 1 variant sharing ancestry with the protective haplotype (**Table 5.3**).

With these variants in hand (a total of 27 on the risk and 20 on the protective haplotypes), we next asked whether any subset of these variants could collectively explain the common haplotype association signals. Starting with the logistic regression model of phenotype ~ haplotype dosage, we greedily selected the individual variant which, when added to the model, most reduced the effect size remaining at the haplotype. We repeated this procedure (adding more and more variants to the regression) until the effect size remaining at the common haplotype could no longer be reduced. Under the hypothesis where low frequency causal variants are responsible for a 'synthetic' common haplotype association, the low frequency variants should collectively be *better* predictors of phenotype than the common haplotype (which is an imperfect proxy), and no residual signal should remain at the common haplotype.

In the case of the risk haplotype, this analysis resulted in the addition of 10 greedily selected low frequency variants to the model (**Figure 5.6c-d**). The risk haplotype, however, still retained a significant (p=2e-03) effect size of OR=1.24 (reduced from 1.33), indicating that low frequency variants are not sufficient to explain the common haplotype's signal. Furthermore, a model including the 10 low frequency variants is not a statistically better predictor of phenotype than dosage of the

risk haplotype alone (log-likelihood ratio test p=0.19). Results for the protective haplotype were similar; 9 low frequency variants were added to the model, but despite their inclusion the haplotype retained an effect size of OR=0.82 (**Figure 5.7d**). Again, the low frequency variants were not a better predictor of phenotype than the common protective haplotype dosage alone.

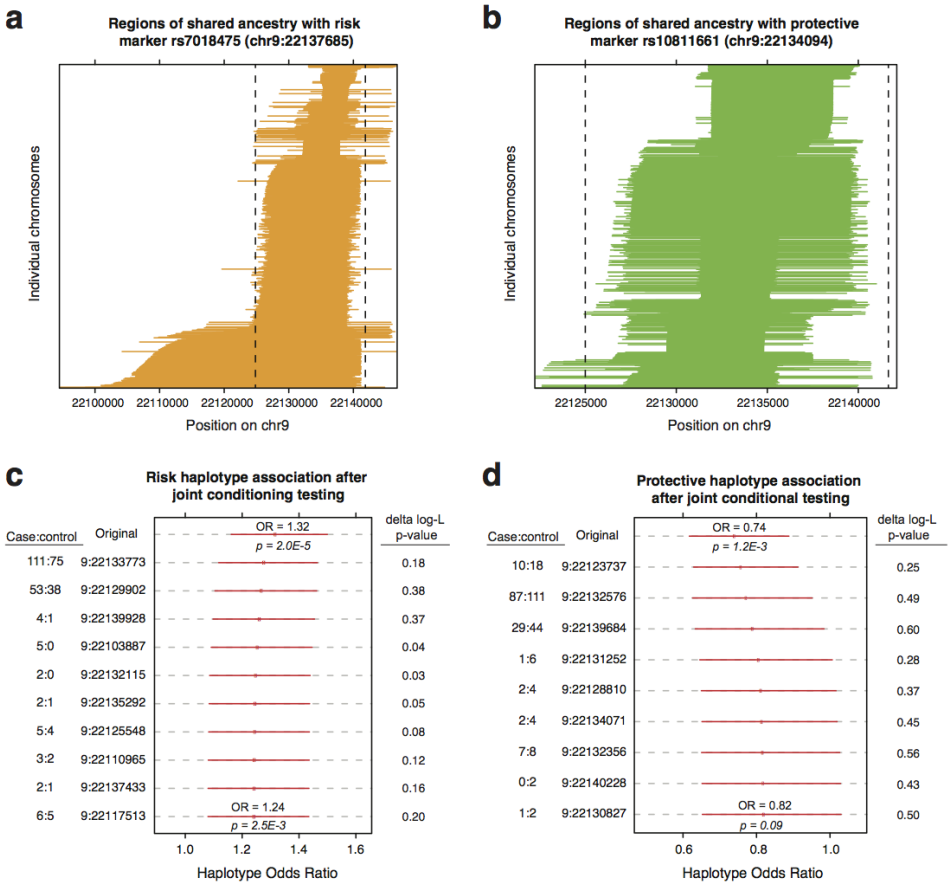


Figure 5.6: Testing for synthetic associations at the chr9p21 locus in the sequenced panel

Ancestral recombination graphs were constructed across the chr9p21 locus. Each graph identifies chromosomal segments (in each individual sample) that share ancestry with a common marker allele. An example is shown **(a)** for a marker of the risk haplotype (chr9:22137685, or rs7018475) and **(b)** for a marker of the protective haplotype (chr9:22134094, or rs10811661). Each horizontal line represents an individual chromosome carrying the minor allele of the marker SNP. Most chromosomes are inferred to have undergone recombination near the known recombination hotspots (dotted black lines), but some samples have longer-range haplotypes. Low frequency and rare variants in regions of shared ancestry were greedily selected for joint association testing along with dosage of the common haplotype, to minimize the residual effect remaining at the common haplotype (**c**,**d**). Variants were added until the common haplotype effect size could not be further reduced. To left of each plot is the added variant's observed case:control counts. Each red line shows the effect size of the haplotype (95% confidence interval shown) in a joint regression model, after cumulative addition of the low frequency and rare variants. In both cases, joint regression models do not have statistically greater explanatory power than the original model with only the common haplotype dosage (p-values for delta log-likelihoods at each step shown at right; not significant).

While the above analyses demonstrate that low frequency variants at 9p21 are unlikely to be driving synthetic common variant associations, it is nonetheless possible that these variants have effects on T2D that are *independent* of the previously described GWAS signals. To test this hypothesis, we (a) used group-based tests of association to evaluate whether rare variants at 9p21 collectively show case-control skew, and (b) conducted joint logistic regression to estimate bounds on the total independent contribution of these variants to T2D heritability.

We ran several group-based tests of association, including those identified in Chapter 4 as having highest power (KBAC and SKAT-O), across all the protein-coding genes in the region as well as across the entire ~10kb region. Across all methods and different frequency thresholds (MAF<0.1%, 1%, or 5%), we observed no significant aggregate association signals for variants across the 9p21 locus (data not shown). The singleton class in particular also showed no case-control skew (27 singletons observed in cases, as compared to 25 singletons observed in controls). It is of course possible (indeed likely) that most variants across the ~10kb locus are neutral, and that only a subset of 'functional' variants would show an aggregate effect; to partially address this, we restricted to only those variants falling within ENCODE-annotated peaks of DNase activity or histone modification, but still observed no significant effects.

The absence of a significant aggregate association signal does not alone exclude the possibility that rare variants at this locus could have causal effects on T2D. Power in this sample to detect effects at very rare alleles or effects of small size (especially if some alleles have effects in the risk direction, while others are protective) is limited (Chapter 4). Moreover, the functional annotations we used above are imprecise; they do not reflect the likelihood that two alleles at a variant site will have *differential* activity, and certainly do not predict whether such activity differences would impact risk of T2D.

Given these limitations, we are evaluating (work in progress) whether the observed sequencing data can enable us to place *bounds* on the total contribution of low frequency variants at 9p21 to T2D heritability, independent of the common signals. Due to linkage disequilibrium, the

contributions of individual low frequency variants cannot simply be added together; this would result in over-estimation. The combined effect of two causal variants on the same haplotype, for example, would be measured at each of the two variants; and conversely, if only one variant is causal, the other variant would also show the same effect. To account for this, we conducted joint logistic regression of phenotype against the full set of low frequency variants (excluding singletons and excluding one of every pair of variants in high LD with each other). We included in this analysis the top common risk and protective variants, so as to measure effects at low frequency variants *independent* of the common signals. The effect sizes estimated in this joint regression model (which is of course over-fit) suggest that rare and low frequency variants at 9p21 could *potentially* contribute ~2% of additional disease heritability in addition to the ~2% explained by common GWAS signals (**Figure 5.7**). No individual rare variant across the locus likely explains more variance than either of the common risk or protective signals. The development of methods to place bounds on these estimates (given large standard errors for rare variant effects), is in progress.

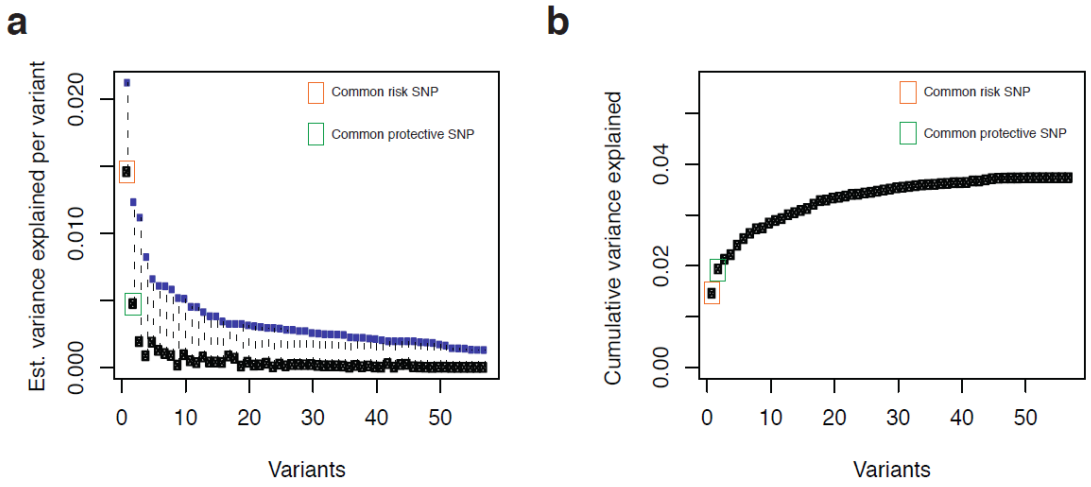


Figure 5.7: Quantifying the contribution of rare variants, independent of common signals at chr9p21

Joint logistic regression was performed, including the dosage of common risk and protective variants as well as the dosage of all (LD-pruned, seen >1x) rare and low frequency variants. The estimated phenotypic variance explained by each variant was calculated by converting odds ratios to relative risks, and then converting this effect to the additive liability scale. The point estimate as well as the 95% upper bound for each individual variant is shown in **(a)**; cumulative variance explained by all variants (using point estimates of their observed effect sizes) is shown in **(b)**.

Taken together, these analyses suggest that the empirical sequencing data at 9p21 is most consistent with a model in which some subset of common variants (specifically, one or more of 2

variants on the risk haplotype, and 13 variants on the protective haplotype) are causal. To further refine this model, we imputed all common variants at the locus into a large panel of ~30K densely genotyped (via the Metabochip array[19,20]) case-control samples. We confirmed that imputation quality was high for these common variants (data not shown), and performed association analysis in each cohort followed by inverse weighted meta-analysis. From meta-analysis association results, we estimated the posterior probability that each variant is causally responsible for the haplotype association signal. This resulted in a clear single signal on the risk haplotype: the 95% credible set for the risk haplotype includes only one variant, rs12555274 (**Table 5.4**). For the protective signal, the 95% credible set includes 5 variants (rs10811662, rs10965250, rs10811661, rs10811660, rs1333051). The top 4 signals on this haplotype are indistinguishable in this cohort; each has a 20-25% probability of being causal.

Table 5.4: Results of association meta-analysis across >30K imputation case-control samples.

All variants identified in sequencing across the chr9p21 locus were imputed into >30K densely genotyped (via the Metabochip) case-control European T2D cohorts. Inverse variance-weighted meta-analysis was performed; effects, standard errors, and p-values observed for all candidate common variants tagging the risk and protective haplotypes are shown above. Posterior probabilities were computed in a Bayesian framework, assuming a fixed prior on the beta for each variant and assuming that a single variant is causal on each haplotype. Variants above the red line are within the 95% credible set of candidate causal variants.

Candidate common risk variants

		Meta-analysis association results			
Variant	ID	se	beta	p-value	Posterior Probability
9:22136440	rs12555274	0.018	-0.1399	7.3E-15	0.998
9:22137685	rs7018475	0.0177	0.1228	4.3E-12	0.002

Candidate common protective variants

		Meta-analysis association results			
Variant	ID	se	beta	p-value	Posterior Probability
9:22134253	rs10811662	0.0214	0.1782	8.6E-17	0.277
9:22133284	rs10965250	0.0214	0.1779	1.1E-16	0.247
9:22134094	rs10811661	0.0214	-0.1774	1.1E-16	0.204
9:22134068	rs10811660	0.0215	0.178	1.1E-16	0.187
9:22136489	rs1333051	0.0236	-0.191	5.3E-16	0.042
9:22132076	rs2383208	0.0212	-0.1698	1.1E-15	0.021
9:22132698	rs10965246	0.0212	-0.1673	3.0E-15	0.008
9:22132729	rs10965247	0.0212	-0.1673	3.0E-15	0.008
9:22132878	rs10965248	0.0212	-0.1659	4.7E-15	0.005
9:22129579	rs7020996	0.0239	0.1665	3.6E-12	0.000
9:22130515	rs10965245	0.028	0.1544	3.6E-08	0.000
9:22130065	rs10965243	0.0273	-0.1416	2.1E-07	0.000
9:22128180	rs12379111	0.0268	-0.127	2.2E-06	0.000

This study systematically tests a wide range of hypotheses about the causal architecture at the T2D-associated chr9p21 locus using sequencing data generated in a large sample of 2,657 European individuals. These data firmly exclude the possibility that rare or low frequency variants

could be driving the observed GWAS signals. Models in which only two common variants (one risk and protective) are causal are sufficient to explain the three-tiered haplotype association, and we find that the 95% credible sets for causal variants on the risk and protective haplotypes include only 1 and 5 common variants, respectively. The data cannot exclude an independent role for rare and low frequency variation, but suggest that these variants could at most explain a limited portion of T2D heritability. Much larger sample sizes would be required to test the rare and low frequency variants for independent association to risk of T2D.

References

1. Hirschhorn, J.N. & Daly, M.J. Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics* **6**, 95-108 (2005).
2. Dickson, S.P., Wang, K., Krantz, I., Hakonarson, H. & Goldstein, D.B. Rare variants create synthetic genome-wide associations. *PLoS Biology* **8**, e1000294 (2010).
3. Anderson, C. a, Soranzo, N., Zeggini, E. & Barrett, J.C. Synthetic associations are unlikely to account for many common disease genome-wide association signals. *PLoS Biology* **9**, e1000580 (2011).
4. Wray, N.R., Purcell, S.M. & Visscher, P.M. Synthetic associations created by rare variants do not explain most GWAS results. *PLoS Biology* **9**, e1000579 (2011).
5. Goldstein, D.B. The importance of synthetic associations will only be resolved empirically. *PLoS Biology* **9**, e1001008 (2011).
6. Wang, K. *et al.* Interpretation of association signals and identification of causal variants from genome-wide association studies. *American Journal of Human Genetics* **86**, 730-42 (2010).
7. Cirulli, E.T. & Goldstein, D.B. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature Reviews Genetics* **11**, 415-25 (2010).
8. McClellan, J. & King, M.-C. Genetic heterogeneity in human disease. *Cell* **141**, 210-7 (2010).
9. Zeggini, E. *et al.* Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science (New York, N.Y.)* **316**, 1336-41 (2007).
10. Saxena, R. *et al.* Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science (New York, N.Y.)* **316**, 1331-6 (2007).
11. Scott, L.J. *et al.* A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science (New York, N.Y.)* **316**, 1341-5 (2007).
12. Kathiresan, S. *et al.* Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. *Nature Genetics* **41**, 334-41 (2009).
13. Sherborne, A.L. *et al.* Variation in CDKN2A at 9p21.3 influences childhood acute lymphoblastic leukemia risk. *Nature Genetics* **42**, 492-4 (2010).
14. Frazer, K. a, Murray, S.S., Schork, N.J. & Topol, E.J. Human genetic variation and its contribution to complex traits. *Nature Reviews Genetics* **10**, 241-51 (2009).
15. The Wellcome Trust Case Control Consortium Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661-78 (2007).
16. Shea, J. *et al.* Comparing strategies to fine-map the association of common SNPs at chromosome 9p21 with type 2 diabetes and myocardial infarction. *Nature Genetics* **43**, 801-5 (2011).
17. Maller, J.B. *et al.* Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nature Genetics* (2012).doi:10.1038/ng.2435
18. Rasmussen, M.D. & Kellis, M. Unified modeling of gene duplication, loss, and coalescence using a locus tree. *Genome Research* **22**, 755-65 (2012).
19. Voight, B.F. *et al.* The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS Genetics* **8**, e1002793 (2012).

20. Morris, A.P. *et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature Genetics* **44**, (2012).

Chapter 6

Insights into allelic architecture across 10 T2D-associated GWAS loci

In this chapter, we apply the fine-mapping methods developed and described in Chapter 5 to the study of ten other type 2 diabetes (T2D) GWAS loci. As in Chapter 5, the majority of analysis was performed using data from a low-pass whole-genome sequencing study of ~2800 unrelated European T2D cases and controls sampled from phenotypic extremes. Variants discovered in this panel were imputed into ~45K additional European samples that have been genotyped on GWAS arrays, and association to T2D across cohorts was combined using inverse-variance weighted meta-analysis (performed by GoT2D Consortium collaborators at the University of Michigan and Oxford). For the majority of T2D GWAS loci, this dataset provides – to date – the most complete catalog of variants that have been genotyped in disease samples and tested for association to T2D.

Somewhat unlike prior chapters – which represent work targeted for publication – this chapter is written primarily to guide future genetic and especially experimental follow-up studies of T2D loci (though analyses presented below will be included in a GoT2D Consortium manuscript currently under preparation). In some cases, the analysis methods deployed are not necessarily novel, but we have endeavored to provide the full set of information (e.g. complete lists of all candidate causal variants, common and rare, observed in the sequencing data) required to design and prioritize experimental studies across a locus. Although there does exist a literature[1,2] (largely restricted to common variants, however) describing fine-mapping of T2D GWAS loci, the data required for follow-up are often not readily available in accessible formats from these publications.

We selected ten T2D loci for fine-mapping on the basis of the strength of association observed across the locus in the GoT2D sequencing dataset, as well as the degree of interest within our lab in functional interrogation of the locus. **Table 6.1** lists these loci, and describes the association signal we observe in GoT2D samples (n=2,657 case-control individuals after quality control) at the previously reported GWAS tag SNP.

Table 6.1: List of T2D GWAS loci studied in this chapter, and association signal observed at the previously reported GWAS tag SNP in GoT2D sequenced panel (n=2,657 samples).

Previously reported T2D GWAS Locus						*GWAS Tag SNP Signal in Metabochip study (Morris et al 2012)*				*Association signal at GWAS SNP in GoT2D WGS panel (n=2,657)*			
Locus Number	Tag SNP rsID	chr:pos (hg19)	Locus interval chr:start-end (hg19)*	Size of LD block (kb)	Closest genes	MAF	Risk allele	OR (risk allele)	p-value	MAF	Allele	OR	p-value
#1	rs7903146	10:114758349	chr10:114545000-114942000	397	TCF7L2	0.27	T	1.39 (1.35-1.42)	1E-139	0.27	T	1.75	2.8E-18
#2	rs11063069	12:4374373	chr12:4314200-4407600	93	CCND2	0.21	G	1.08 (1.05-1.11)	3.3E-07	0.22	G	1.20	7.0E-03
#3	rs1531343	12:66174894	chr12:66158000-66399000	241	HMGA2	[0.07-0.13]	C	1.15 (1.09-1.22)	4.9E-07	0.09	C	1.37	1.3E-03
#4	rs4430796	17:36098040	chr17:36020000-36136000	116	HNF1B	[0.50-0.66]	G	1.13 (1.07-1.19)	2.4E-06	0.43	G	1.19	2.3E-03
#5	rs4402960	3:185511687	chr3:184400000-185750000	1350	IGF2BP2	0.33	T	1.13 (1.10-1.16)	2.4E-23	0.32	T	1.25	1.7E-04
#6	rs1801282	3:12393125	chr3:12023800-12863000	839	PPARG	0.86	C	1.13 (1.09-1.17)	1.1E-12	0.15	G	0.82	9.7E-03
#7	rs7756992	6:20679709	chr6:20413200-21294500	881	CDKAL1	0.29	G	1.17 (1.14-1.20)	7.0E-35	0.31	G	1.12	0.05
#8	rs849134	7:28196222	chr7:28030000-28264000	234	JAZF1	[0.47-0.50]	A	1.12 (1.08-1.16)	3.2E-10	0.48	G	0.97	0.56
#9	rs5215	11:17408630	chr11:16995000-17435000	440	KCNJ11	0.41	C	1.07 (1.05-1.10)	8.5E-10	0.44	C	1.14	0.02
#10	rs231362	11:2691471	chr11:2635000-2865000	230	KCNQ1	0.52	G	1.08 (1.05-1.11)	1.7E-09	0.47	A	0.93	0.19

Within this chapter, we provide detailed descriptions of the signal at the first three loci (TCF7L2, CCND2, and HMGA2); these examples (alongside chr9p21) highlight the diversity of genetic data across T2D GWAS loci. For the other seven loci, we provide only high-level summaries to give the reader a sense of which loci might be most interesting and also tractable to study. At the three loci described in depth, we organize findings in the below format, in order to ask and answer the following questions:

A) Characterization of the previously reported GWAS tag SNP and its LD partners
   a. How many variants are in LD with the tag SNP (a measure of haplotype complexity)?
   b. Are any of these tag SNPs protein-coding (attractive for follow-up)?

B) Regional association signal(s) observed in the GoT2D sequenced panel
   a. Do any low frequency variants have signal exceeding that of the GWAS tag SNP?
   b. How many independent signals are present across the locus?

C) Could low frequency or rare variants at the locus be producing synthetic associations?
   a. Can an individual low frequency variant drive the GWAS signal?
   b. Could a group of low frequency or rare variants be responsible, in aggregate?

D) What is the role of rare variation, independent of the common GWAS signal?

   a. Is there a burden of T2D-associated variants across protein-coding genes?
   b. Which individual low frequency or rare variants show strongest association (e.g. for follow-up genotyping or experimental testing)?

E) Summary of findings at the locus and recommendations for follow-up

Locus #1: TCF7L2

A) ***Characterization of the previously reported GWAS tag SNP and its LD partners***

The GWAS tag SNP at this locus is **rs7903146** (chr10:114,758,349; MAF=0.27; coordinates in hg19 throughout this chapter). There is robust association to T2D observed at this SNP in the GoT2D panel (OR = 1.75; p= 2.7E-18; it is the third most associated variant in this dataset). None of the top T2D-associated variants are rare or of low frequency (MAF<5%); however, one small insertion (G>GCT at chr10:114782581 is among the top few signals (**Figure 6.1b**) and is in strong LD with the GWAS tag SNP (r^2 = 0.87). There are **6 common variants in strong LD** (r^2 >= 0.8) with the GWAS tag SNP and **32 variants in modest LD** (r^2 >= 0.5); none of these are protein-coding, and all fall within an intron of TCF7L2. From this set, the top ten variants are listed below (sorted by association p-value). The 6 variants in strong LD with the tag SNP show strongest signal in both the sequenced panel as well as in imputation analysis.

Table 6.2: Top common variants in LD with GWAS tag SNP (rs7903146)

	Variant	MAF	r2 with GWAS tag	OR	P-value	Annotation	Imputation p-value (~50K samples)
	10:114754784	0.2695	0.99	1.757	1.54E-18	INTRONIC (TCF7L2)	1.60E-70
	10:114754071	0.2697	0.99	1.753	1.97E-18	INTRONIC (TCF7L2)	3.95E-70
GWAS tag SNP	10:114758349	0.2678	1.00	1.751	2.80E-18	INTRONIC (TCF7L2)	1.87E-71
	10:114782581:indel	0.2753105	0.87	1.725	2.6E-17	INTRONIC (TCF7L2)	not imputed
	10:114785424	0.2719	0.88	1.714	4.27E-17	INTRONIC (TCF7L2)	1.33E-66
	10:114774433	0.2734	0.88	1.712	4.47E-17	INTRONIC (TCF7L2)	2.93E-67
	10:114808902	0.2522	0.75	1.695	1.86E-15	INTRONIC (TCF7L2)	1.65E-60
	10:114791490	0.2505	0.77	1.688	3.19E-15	INTRONIC (TCF7L2)	2.12E-61
	10:114798893	0.2503	0.77	1.685	3.92E-15	INTRONIC (TCF7L2)	5.39E-61
	10:114788815	0.2499	0.77	1.683	4.20E-15	INTRONIC (TCF7L2)	4.65E-61

B) ***Regional association signal(s) observed in the GoT2D sequenced panel***

Across both a 5Mb region surrounding this SNP (chr10:112250000-117250000) as well as a narrower interval defined by recombination hotspots around the tag SNP (chr10:114545000-114942000), the previously described GWAS signal is the only clear association observed (**Figure 6.1a-b**). After conditioning on the top variant listed in the table above (chr10:114754784), no significant residual signal remains within the GWAS locus (**Figure 6.1c**), suggesting that there is

only one independent common variant T2D association signal within the GWAS locus (as defined by recombination hotspots). Across the larger 5Mb interval, many low frequency variants show modest association signals (p-values in the 0.001 – 0.0001 range), but most of these signals are no longer present after conditioning on the GWAS tag SNP. A few such rare variants retain independent association after conditioning; these are discussed in section D.

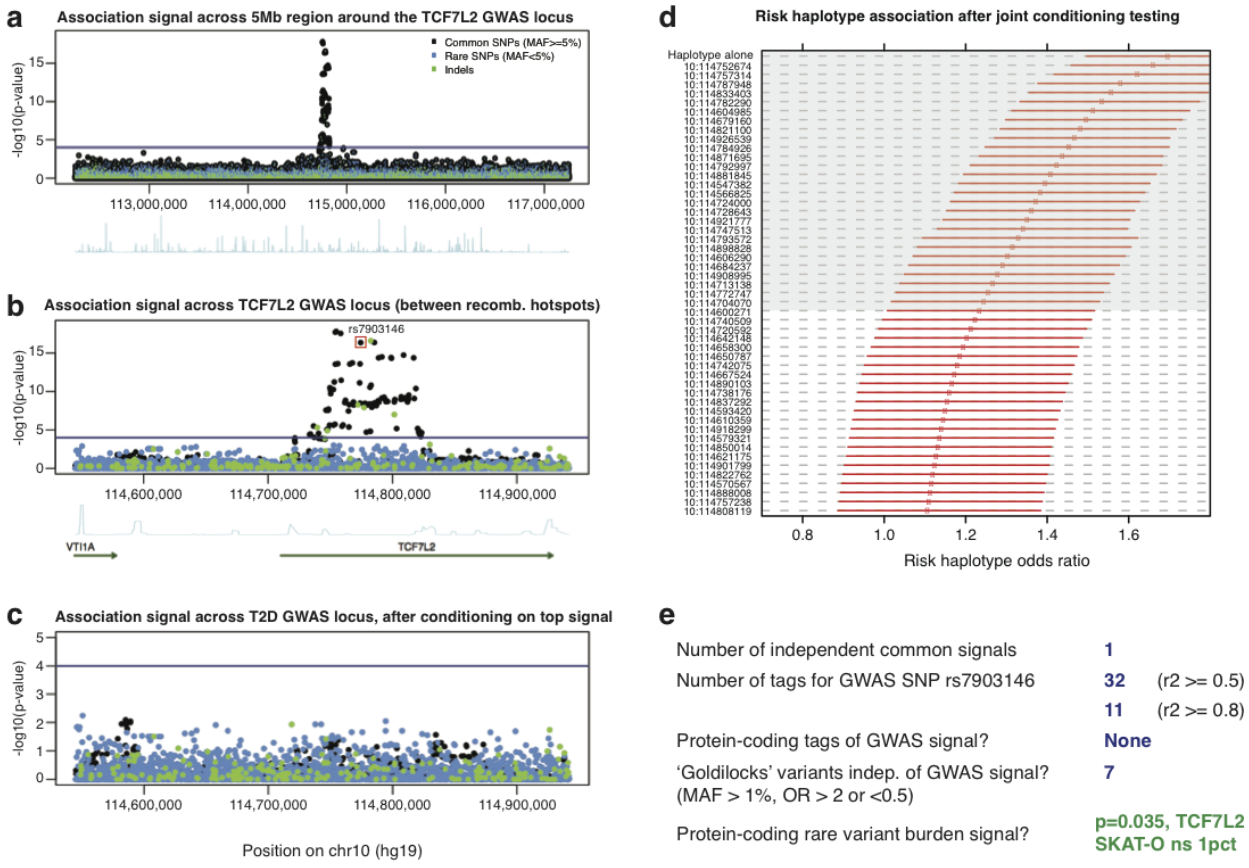


Figure 6.1: Fine-mapping of TCF7L2 locus in GoT2D (n=2,657) whole genome sequencing panel

C) *Could low frequency / rare variants at the locus be producing synthetic associations?*

We identify 455 rare and low frequency (MAC > 1; MAF < 5%) variants segregating on the background of the GWAS T2D-associated common haplotype (and located within the interval defined by recombination hotspots). No single variant among these can explain the common haplotype's association; over 50 variants would be required to reduce the effect size (OR) at the

haplotype to below 1.10 (**Figure 6.1d**). The likelihood of a genetic model with only these 50 greedily selected rare and low-frequency variants (relative to the likelihood of a model in which only the common haplotype explains disease risk) is 0.002 (as assessed by AIC); these variants do not have greater explanatory power than the common signal alone. Thus, at TCF7L2, rare and low frequency variants are extremely unlikely to explain the observed common variant GWAS signal.

D) ***What is the role of rare variation, independent of the common GWAS signal?***

Across the TCF7L2 GWAS interval, there are no 'Goldilocks' alleles (defined here as having MAF > 1%, OR > 2 or OR < 0.5, p-value < 0.01) that retain association (p<0.01) after conditioning on the GWAS tag SNP. However, across the larger 5Mb interval, there are a handful of such variants (none are protein-coding):

Table 6.3: Potential Goldilocks' low frequency variants across the broader 5Mb TCF7L2 interval

Variant	MAF	Case Obs	Control Obs	OR	P-value	Conditional P-val	Protein-coding?	Imputation p-value (~50K samples)
10:114962203	1.8%	66	29	2.32	1.74E-04	3.35E-04	No	0.778
10:112942235	2.1%	38	74	0.49	5.13E-04	6.20E-04	No	0.466
10:112286195:indel	1.3%	48	21	2.39	1.10E-03	3.55E-04	No	-
10:115002009	1.4%	51	23	2.25	1.23E-03	2.10E-03	No	0.742
10:115698102	1.0%	15	39	0.38	1.55E-03	9.64E-04	No	0.285
10:115557050	1.3%	23	45	0.48	5.63E-03	1.76E-03	No	0.178
10:115797837	1.1%	18	38	0.46	6.95E-03	9.23E-03	No	0.247

As seen above, none of these variants shows any association after imputation meta-analysis It should be noted that imputation quality for variants of frequency <2% is highly variable, and this has not been evaluated for the variants above. Nonetheless, given the absence of replication in the imputation cohorts, these low frequency variants should be further evaluated in ongoing genetic studies prior to experimental interrogation (especially since they are all rather far away from both the original T2D GWAS signal and the TCF7L2 transcript, and there is thus little prior biological reason to nominate these variants as causally associated with risk of T2D).

Across the GWAS interval, there are two protein-coding genes (as seen in **Figure 6.1b**). We applied the gene-based association test SKAT-O to non-synonymous (MAF<1%) as well as loss-of-function variants (LOF; includes nonsense and frame-shift) identified within these genes. No signal

was observed at VTI1A; at TCF7L2, however we observe nominal association across non-synonymous variants (SKAT-O p-value = 0.035; p-value after including GWAS tag SNP as a covariate = 0.032). This (weak) signal is driven mainly by a single variant (top listed below):

Table 6.4: Rare non-synonymous (MAF<1%) variants in TCF7L2 gene driving burden signal

Variant	Case Obs	Control Obs	OR	P-value	Annotation; PolyPhen Category
10:114925406	7	19	0.39	0.03	Missense (p.P495R); possibly damaging
10:114849271	7	2	4.28	0.07	Missense (p.Q199R); benign
10:114711358	3	3	1.11	0.90	Missense (p.A125T); benign
10:114711262	1	0	-	-	Missense (p.G93R); probably damaging
10:114903696	1	0	-	-	Missense (p.P234S); possibly damaging
10:114903756	1	0	-	-	Missense (p.P254S); benign
10:114925622	1	0	-	-	Missense (p.A567V); benign
10:114849163	0	1	-	-	Missense (p.P163R); benign
10:114910857	0	1	-	-	Missense (p.S326G); possibly damaging
10:114925441	0	1	-	-	Missense (p.A507T); benign
Total	21	27			

No LOF variants were identified in the TCF7L2 gene.

E) ***Summary of findings***

At the TCF7L2 locus, we do not observe any strong low frequency or rare variant associations that are independent of the GWAS signal (though we cannot exclude the possibility of *any* rare variant effects; some of the potentially interesting such variants are listed in **Table 6.3** and **6.4**). The previously known genetic association to T2D at the TCF7L2 locus appears to be driven by a single common haplotype (on which one or more common variants may be causal). Because >50 rare and low frequency variants would be required to explain the large effect at this haplotype, such a model is extremely unlikely. The set of most likely candidate causal variants is comprised of 6 **common** variants (5 SNPs and one small insertion) that are in tight LD with the previously reported tag GWAS SNP. These are the top 6 variants listed in **Table 6.2**.

Locus #2: CCND2

A) ***Characterization of the previously reported GWAS tag SNP and its LD partners***

The previously reported GWAS tag SNP at the CCND2 locus is **rs11063069** (chr12:4374373; MAF = 0.22). Interestingly, this SNP was also reported to have significant evidence for sex-differentiated association to T2D (male OR = 1.12; female OR = 1.04; heterogeneity p=0.013).[1] In the GoT2D panel, this SNP shows modest association signal (OR = 1.20; p= 0.007; **Table 6.1**). It lies on a haplotype of extremely low complexity; there exist only 3 common variants in even modest LD (r^2 >= 0.5) with the GWAS tag SNP:

Table 6.5: Common variants in LD with CCND2 GWAS tag SNP (rs11063069)

	Variant	MAF	r2 with GWAS tag	OR	P-value	Annotation	Imputation p-value (~50K samples)
	12:4362909	0.22	0.93	1.208	4.73E-03	INTRONIC (NON_CODING_GENE)	3.03E-07
	12:4376089	0.22	0.98	1.205	5.18E-03	INTRONIC (NON_CODING_GENE)	3.47E-08
GWAS tag SNP	12:4374373	0.22	1.00	1.197	7.00E-03	REGULATORY REGION	9.05E-08

Each SNP shows similar association to T2D. None of these variants are insertions or deletions; none are protein-coding. All three SNPs are entirely *upstream* of the CCND2 transcript (though two of them are annotated as falling within the intron of a different non-coding transcript).

B) ***Regional association signal(s) observed in the GoT2D sequenced panel***

The regional association plot across the CCND2 locus is striking because a lone (and novel) low frequency variant (12:4384844; rs76895963; MAF = 2.3%; OR = 0.35; p = 5.9e-07; shown in blue in **Figure 6.2b**) shows stronger association to T2D than any common variants. This is the strongest signal not only across an interval defined by recombination hotspots flanking the GWAS tag SNP (chr12:4314200-4407600), but also across a large 5Mb interval (chr12:1860000-6860000) Interestingly, the most-associated common variant observed in the GoT2D data (12:4379831; MAF = 0.23; OR = 0.78; p=0.00017) is actually a variant tagging the common haplotype on which 12:4384844 occurred; 86% of all observations of the low frequency variant 12:4384844 occur in phase with the common variant 12:4379831). Thus, it is likely that the low frequency variant is actually driving weak association at this common SNP (which has very low r2 = 0.08 to the GWAS

tag SNP chr12:4374373). The novel rare SNP is independent of the previously reported GWAS tag SNP (r^2 = 0.007 to chr12:4374373).

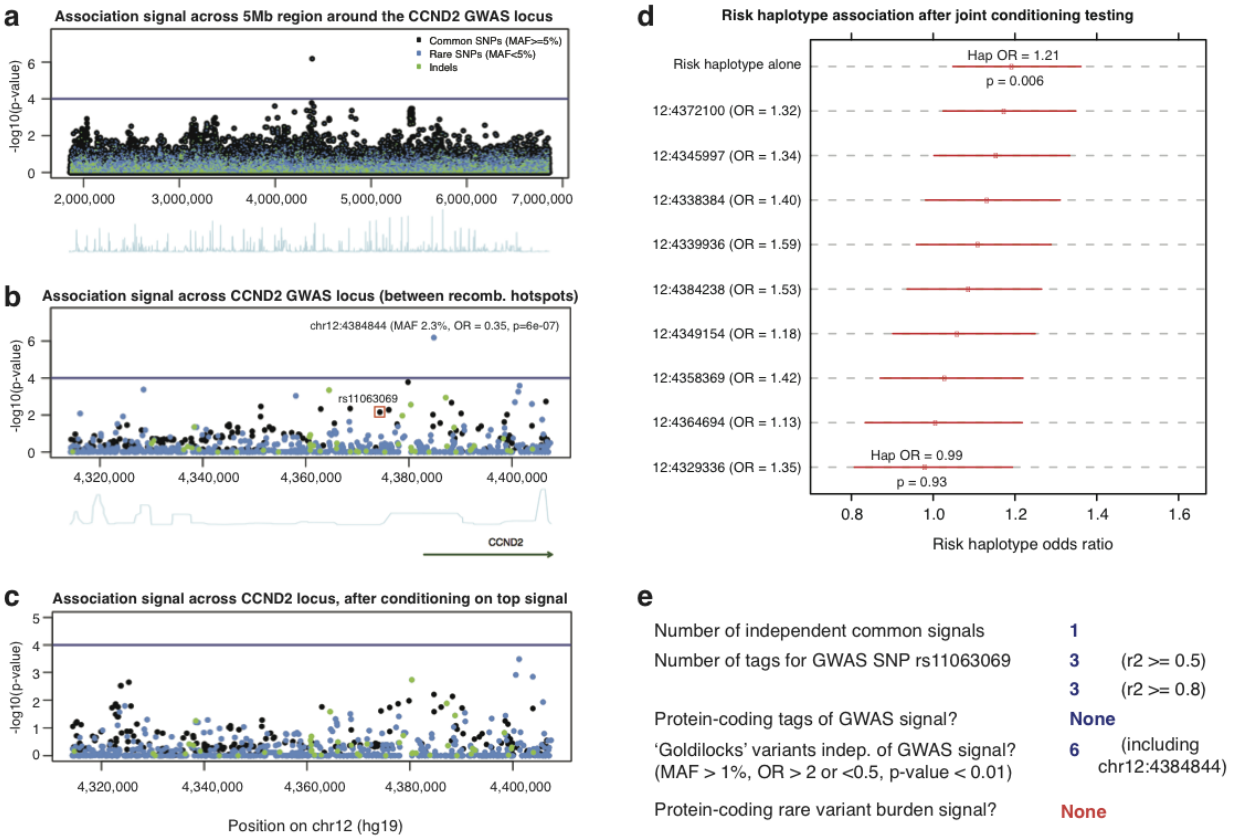


Figure 6.2: Fine-mapping of CCND2 locus in GoT2D (n=2,657) whole genome sequencing panel

After conditioning on the top low frequency signal (12:4384844), no strong independent signals are observed in this dataset (**Figure 6.2c**), though a few rare SNPs retain modest association (p~0.001). Association signal at the GWAS tag SNP 12:4374373 is slightly reduced (OR=1.16, p=0.024), but the bulk of the effect size remains. This suggests that there are indeed (at least) two independent signals at this locus: one common haplotype which is associated with *increased T2D risk* (tagged by 12:4374373), and another rare variant which is associated with *decreased T2D risk*.

Visualization of the haplotype structure at CCND2 (and phasing of the rare variant onto common haplotypes, as described in Chapter 5) further supports the conclusion that these are independent effects (**Figure 6.3**). The diagram also clarifies why signal at the GWAS tag (red) is

slightly reduced after conditioning on the rare variant (blue); a rare protective signal on a *non-risk* common haplotype (which does *not* have the GWAS risk allele) can explain a small fraction of the GWAS signal. If only the rare protective variant (blue) were causal, however, then all common haplotypes *not* carrying it should appear uniformly (and very weakly) associated with increased risk of T2D; instead we observe clear association in the risk direction at only H2 and at no other haplotypes. This observation here (in the GoT2D sequencing data), combined with our prior knowledge that variants tagging H2 are associated to risk of T2D at genome-wide significance levels in GWAS – suggests that indeed there are two independent signals at this locus.

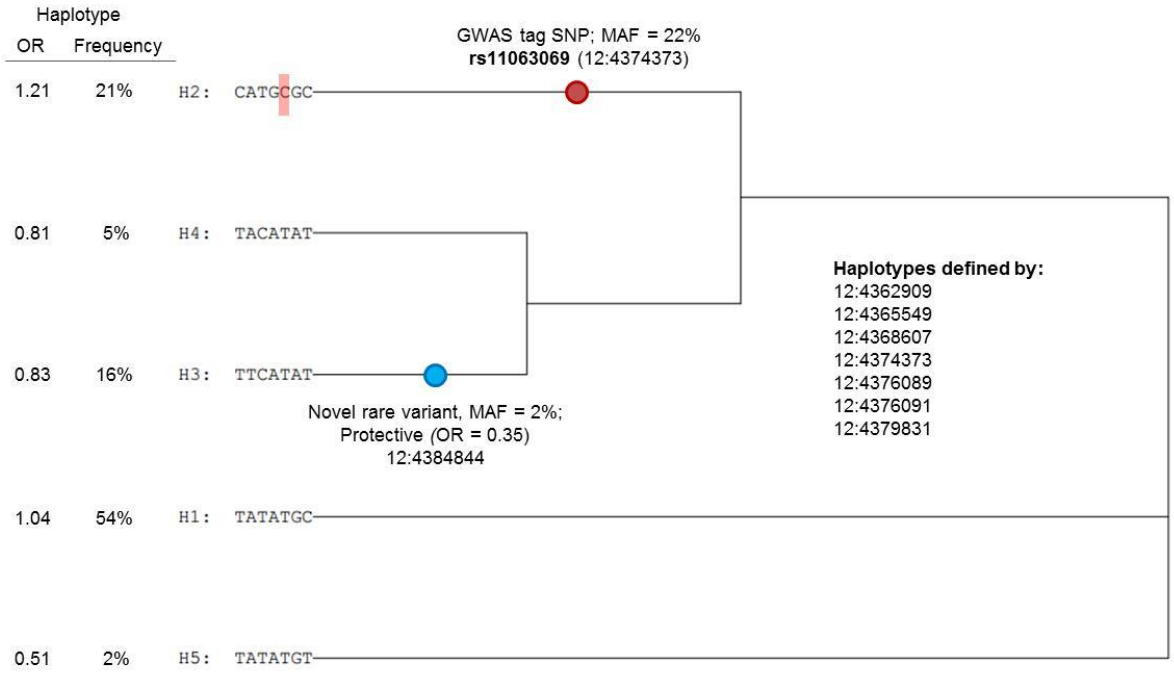


Figure 6.3: Fine-mapping of CCND2 locus in GoT2D (n=2,657) whole genome sequencing panel

C) *Could low frequency / rare variants at the locus be producing synthetic associations?*

As discussed above, the novel low frequency variant (12:4384844) does indeed induce weak 'synthetic' association at common SNPs that tag the haplotype on which it arose. However, these common variants are not in LD with the disease-associated markers reported in T2D GWAS. We

wondered whether the previously known GWAS signal could be explained by low frequency or rare variants.

We identify 85 rare and low frequency (MAC > 1; MAF < 5%) variants segregating on the background of the GWAS T2D-associated common haplotype (on H2; located within the interval defined by recombination hotspots). No single variant among these can explain the entire effect at the common haplotype; however, 9 variants would be sufficient to reduce the haplotype effect size (as measured by OR) to < 1.0 (**Figure 6.2d**). A closer look at these variants (**Table 6.6**) reveals that these are exactly the kind of variants we might expect to potentially produce a synthetic association: their observations almost always occur on the common risk haplotype, and they have larger effect sizes than the common GWAS risk allele.

Table 6.6: Rare / low frequency variants sufficient to explain CCND2 GWAS signal in GoT2D panel

Variant	MAF	MAC	% on risk haplotype	Case Obs	Control Obs	OR	Single variant P-value
12:4372100	2.2%	117	98%	67	50	1.32	0.14
12:4345997	1.7%	91	90%	52	39	1.34	0.17
12:4338384	2.8%	148	50%	85	63	1.40	0.05
12:4339936	0.9%	48	83%	29	19	1.59	0.12
12:4384238	1.1%	57	88%	35	22	1.53	0.13
12:4349154	3.3%	175	94%	96	79	1.18	0.30
12:4358369	1.3%	70	86%	41	29	1.42	0.16
12:4364694	3.5%	187	98%	98	89	1.13	0.42
12:4329336	0.8%	40	80%	23	17	1.35	0.34

The likelihood of a genetic model with only these 9 greedily selected rare and low-frequency variants (relative to the likelihood of a model in which only the common haplotype explains disease risk) is 0.12 (as assessed by AIC). Thus, these variants *do not have greater explanatory power* than the common haplotype alone (though it remains possible that they are causally driving the GWAS association). Certainly, the more parsimonious explanation for the CCND2 GWAS signal is that a single common variant (one of the three variants listed in **Table 6.5**) is causal.

D) ***What is the role of rare variation, independent of the common GWAS signal?***

We looked across both the 5Mb-interval as well as the narrower GWAS interval for putative Goldilocks (MAF > 0.01, OR > 2 or OR < 0.5, p-value < 0.01) variants (**Table 6.7**). As seen above, the novel rare variant 12:4384844 represents the strongest such signal; within the GWAS interval, 3 other variants in modest LD with this variant are also identified (all have weaker signal both in the sequenced panel as well as in the imputed data). It is worth noting that 12:4384844 retains significant association even after imputation meta-analysis in ~50K samples, providing some further evidence for its candidacy as a causal variant; however, the fact that the p-value only improves from 6.5e-07 to 2.8e-07 indicates that the effect size of this variant was much less in the other cohorts examined (cohorts other than the GoT2D sequenced panel).

Table 6.7: Potential Goldilocks' low frequency variants at the CCND2 locus

Variant	MAF	Case Obs	Control Obs	OR	P-value	P-value, conditioned on GWAS tag	P-value, conditioned on 12:4384844	r2 with 12:4384844	Protein-coding?	Imputation p-value (~50K samples)
12:4384844	2.3%	32	89	0.35	6.54E-07	1.74E-06	-	-	No	2.83E-07
12:4401472*	0.8%	8	34	0.23	2.59E-04	3.82E-04	0.14	0.31	No	2.09E-06
12:4401201	3.0%	57	101	0.56	5.48E-04	6.28E-04	3.30E-04	0.001	No	0.78
12:4328521*	1.6%	27	60	0.44	4.22E-04	7.09E-04	0.16	0.22	No	6.87E-04
12:4358078*	1.0%	15	40	0.36	9.21E-04	1.46E-03	0.15	0.18	No	3.56E-04
In the broader 5Mb interval:										
12:5633793	1.4%	50	23	2.09	3.85E-03	3.39E-03	0.004	0.00	No	0.15
12:3276210	1.4%	25	51	0.50	5.19E-03	5.66E-03	0.005	0.00	No	0.44
12:3286771	1.4%	25	51	0.50	5.19E-03	5.66E-03	0.005	0.00	No	0.39

*Occur on the same common haplotype background as the top signal (12:4384844)

A variant at 12:4401201 is included in **Table 6.7** above (despite not meeting our stringent definition of a Goldilocks variant) because it is another low frequency variant that falls within an intron of CCND2; given the signal at 12:4384844, we might be more interested in similar variants, especially if they are independent (as 12:4401201 is; r2 to 12:4384844 = 0.001). However – underscoring the challenges of interpreting association data for putative novel signals in relatively small sample sizes – this variant shows no association after imputation meta-analysis (p=0.78). Low imputation quality is a concern, but this variant has relatively high frequency (MAF=3%). Further

genotyping of this variant could be warranted, but at present the evidence supporting its association to T2D is weak.

Similarly, across the larger 5-Mb interval, we observe three other variants that show moderately strong signal in the GoT2D panel, and retain association after conditioning on the other signals (unsurprisingly, given the distance between these variants and the GWAS interval). However, none of these variants show any association in the imputation meta-analysis.

Across this GWAS interval, CCND2 is the only protein-coding gene (as seen in **Figure 6.2b**). No significant burden of non-synonymous variation (MAF<1%) was observed in this gene (SKAT-O p=0.46). No loss-of-function variants were identified in CCND2.

E) ***Summary of findings***

The data at the CCND2 locus suggest that there are two independent genetic associations to T2D here (one common signal, and one low frequency signal). The common signal is tagged by only 3 common SNPs (listed in **Table 6.5**); one or more of these 3 variants is most likely to causally drive the T2D risk signal observed in GWAS. It is also possible that a collection of (as few as 9) low frequency variants at this locus *could* be driving the GWAS signal (such a set is listed in **Table 6.6**), but this genetic model is less likely than models implicating a common causal variant.

We detect a second independent, low frequency *protective* signal at this locus at the non-coding variant 12:4384844 (rs76895963; MAF=2.3%; OR = 0.35; imputation meta-analysis p=3e-07). Only one other variant (of lower frequency, in nearly perfect D' with 12:4384844) shows signal of comparable strength after large-scale imputation: 12:4401472 (MAF=0.8%; p=2e-06 in meta-analysis). If follow-up genotyping or experimental studies are undertaken, both these variants should be further interrogated.

The CCND2 locus is a particularly attractive one for functional follow-up because of its relatively low complexity: only 5 variants with high likelihood of being causal have been identified (1 of 3 variants tagging the common risk haplotype, and 1 of 2 low frequency variants). A key challenge, of course, is that all these variants lie in non-protein-coding regions of the genome.

Locus #3: HMGA2

A) ***Characterization of the previously reported GWAS tag SNP and its LD partners***

The previously reported GWAS tag SNP at the HMGA2 locus is **rs1531343** (12:66174894; MAF = 0.09). In the GoT2D panel, this SNP has a relatively large effect size (OR = 1.37; p= 0.0013; **Table 6.1**). In stark contrast to the CCND2 locus, this SNP lies on a haplotype of very high complexity; there exist 62 common variants in tight LD (r^2 >= 0.8) with the GWAS tag SNP, and 132 common variants in modest LD (r^2 >= 0.5). Because such a large number of variants show similar association signal in the GoT2D panel (seen as a thick row of points in the Manhattan plot in **Figure 6.4a**), we relied on association signal in the imputation meta-analysis to build a credible set of candidate causal variants. Using a Bayesian analysis (as done in Chapter 5, and as previously described[2]), we find that a 90% credible set of candidate causal variants contains 18 common SNPs (listed in **Table 6.8**); a 95% credible set contains 35 common SNPs. These sets must be interpreted with caution, however, as the incremental differences between variants inside and outside these sets is extremely small (the original tag SNP, for example, is actually *not* included in either of these sets).

Table 6.8: Common SNPs in 90% credible set, ranked by imputation p-value

Variant	MAF	r2 with GWAS tag	OR	P-value	Annotation	Imputation p-value (~50K samples)	Posterior probability
12:66211006	0.089	0.93	1.37	1.4E-03	INTRONIC (NON_CODING_GENE)	4.4E-07	0.102
12:66215292	0.089	0.93	1.36	1.6E-03	UPSTREAM	4.6E-07	0.094
12:66210621	0.088	0.93	1.35	1.9E-03	INTRONIC (NON_CODING_GENE)	5.8E-07	0.079
12:66200749	0.088	0.94	1.36	1.6E-03	INTRONIC (NON_CODING_GENE)	6.1E-07	0.074
12:66208513	0.088	0.93	1.35	1.9E-03	INTRONIC (NON_CODING_GENE)	6.5E-07	0.069
12:66209646	0.088	0.93	1.35	1.9E-03	INTRONIC (NON_CODING_GENE)	6.6E-07	0.069
12:66213258	0.088	0.93	1.35	1.9E-03	UPSTREAM	7.0E-07	0.067
12:66205071	0.089	0.95	1.37	1.4E-03	INTRONIC (NON_CODING_GENE)	7.6E-07	0.058
12:66212318	0.088	0.94	1.34	2.5E-03	INTRONIC (NON_CODING_GENE)	7.8E-07	0.056
12:66199734	0.072	0.76	1.33	8.3E-03	INTRONIC (NON_CODING_GENE)	9.1E-07	0.053
12:66204598	0.088	0.95	1.35	1.9E-03	INTRONIC (NON_CODING_GENE)	9.3E-07	0.049
12:66204696	0.088	0.95	1.35	1.9E-03	INTRONIC (NON_CODING_GENE)	9.7E-07	0.048
12:66203383	0.088	0.95	1.35	1.9E-03	INTRONIC (NON_CODING_GENE)	1.0E-06	0.044
12:66221060	0.094	0.87	1.37	8.0E-04	INTRONIC (HMGA2)	3.6E-06	0.013
12:66194613	0.088	0.96	1.37	1.3E-03	INTRONIC (NON_CODING_GENE)	5.1E-06	0.010
12:66193110	0.088	0.96	1.37	1.3E-03	INTRONIC (NON_CODING_GENE)	6.4E-06	0.008
12:66192667	0.088	0.96	1.36	1.6E-03	INTRONIC (NON_CODING_GENE)	8.4E-06	0.006
12:66194243	0.104	0.80	1.29	4.7E-03	INTRONIC (NON_CODING_GENE)	1.1E-05	0.005

Additionally, seven common insertion/deletion variants (**Table 6.9**) are also in tight LD with the tag SNP and show association signal in the GoT2D sequenced panel of comparable strength to the above SNPs. Because these variants have not yet been imputed, they were not included in the Bayesian credible set analysis, but they must be considered in any future studies aimed at identifying causal variants at this locus.

Table 6.9: Top common insertions/deletions in LD with tag SNP (rs1531343)

Variant	MAF	r2 with GWAS tag	OR	P-value
12:66173114:indel	0.089	0.99	1.37	1.1E-03
12:66176445:indel	0.089	1.00	1.37	1.3E-03
12:66199779:indel	0.089	0.94	1.37	1.3E-03
12:66184400:indel	0.091	0.98	1.36	1.5E-03
12:66176268:indel	0.089	0.99	1.36	1.6E-03
12:66186277:indel	0.087	0.96	1.36	1.7E-03
12:66171677:indel	0.100	0.88	1.29	6.2E-03

None of the variants in modest LD with the GWAS tag SNP are protein-coding.

B) ***Regional association signal(s) observed in the GoT2D sequenced panel***

Across a 5Mb region surrounding the tag SNP (chr12:63674894-68674894), the previously described GWAS signal appears to be one of many comparable signals (with association p-value ~ 0.001); these other peaks could be further evaluated in the imputation meta-analysis to determine whether they likely reflect true association to T2D or noise in this relatively small sample size. Here, we zoom into the narrower interval defined by recombination hotspots around the GWAS tag SNP (chr12:66158000-66399000) to characterize the signal that has been previously replicated in very large sample sizes (**Figure 6.4b**). In this interval, we observe a think cluster of variants – the SNPs and insertion/deletion variants listed in section A – that all show comparable association signal. No rare or low frequency variants appear to be part of this set. The top signal in this dataset (12:66221060) is in tight LD with the prior GWAS tag SNP ($r^2 = 0.87$).

After conditioning on this top signal, no residual signal remains across the locus, indicating that there is only one independent signal at the HMGA2 locus (that is detectable in this sample).

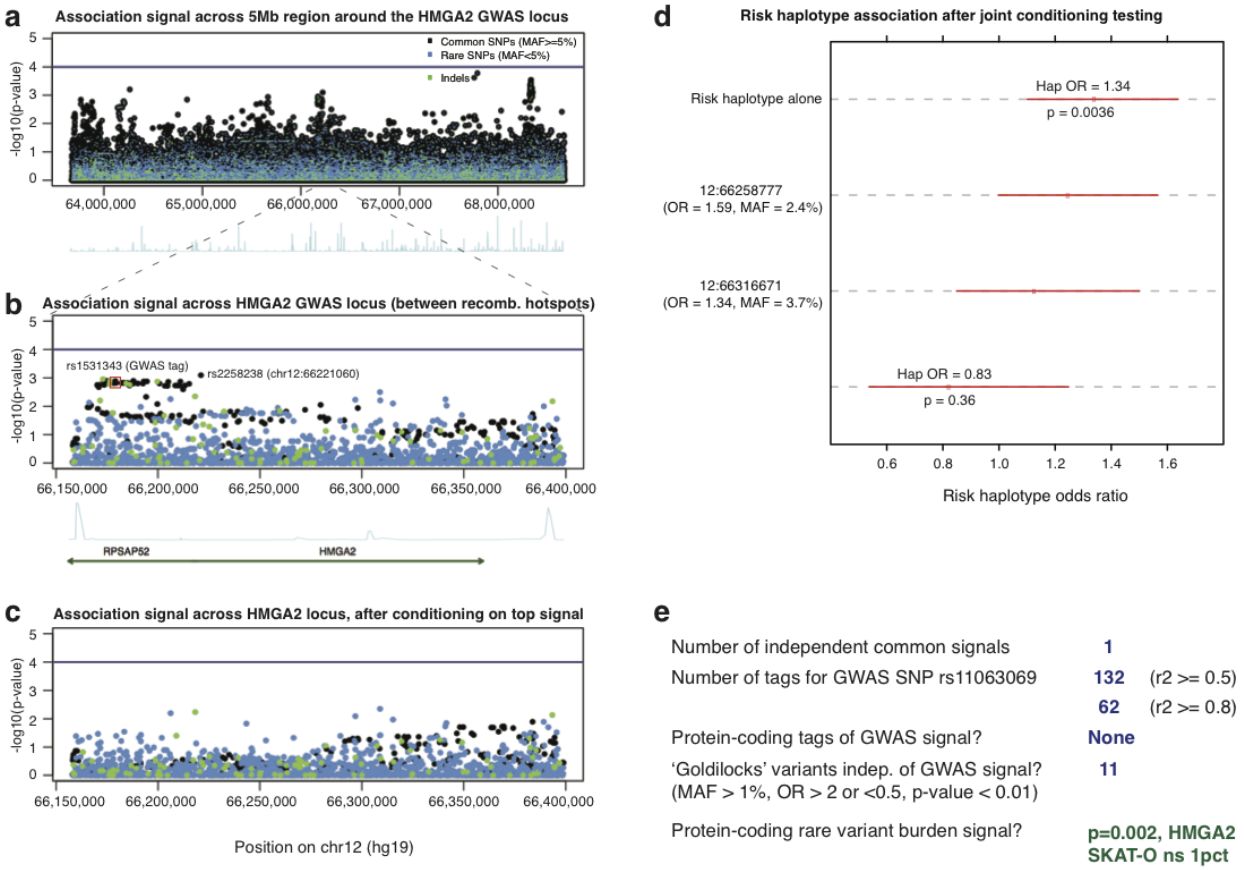


Figure 6.4: Fine-mapping of HMGA2 locus in GoT2D (n=2,657) whole genome sequencing panel

C) *Could low frequency / rare variants at the locus be producing synthetic associations?*

We identify 142 rare and low frequency variants segregating on the GWAS risk haplotypes at the HMGA2 locus. In part because the risk haplotype (and the GWAS tag SNP) at this locus is of relatively lower frequency (~9%), we find that only three low frequency variants would be sufficient to explain the GWAS signal. There are actually 44 low frequency variants on the risk haplotype with MAF>2% and OR>1.3; any combination of three to four such variants would be sufficient to explain the GWAS signal. Just one example of three such variants is shown in **Figure 6.4d** and listed in **Table 6.10** (these should not be interpreted as the only three low frequency variants with this property). In this case (unlike what we observed at CCND2), a genetic model with only these three low frequency variants actually has *greater likelihood* of explaining the observed data as compared to a model with the common risk haplotype alone (~2x more likely; based on AIC). This certainly

does not rule out genetic models with common causal variants, but low frequency models at this locus must be considered. It is important to keep in mind that in the scenario where a common variant was causal, it would be entirely *expected* for low frequency variants on the haplotype to show association in the same direction as the haplotype. It just so happens that at this locus, a large number of such low frequency variants are segregating in the population, making the alternative low frequency model plausible.

Table 6.10: Example of low frequency variants sufficient to explain HMGA2 GWAS signal in GoT2D

Variant	MAF	MAC	% on risk haplotype	Case Obs	Control Obs	OR	P-value in GoT2D	Imputation p-value (~50K samples)
12:66258777	2.4%	130	89%	79	51	1.59	0.01	4.23E-04
12:66316671	3.7%	198	77%	113	85	1.34	0.05	9.51E-03
12:66180277	2.7%	143	85%	82	61	1.41	0.05	0.30

It is also worth noting that association signal at some of these candidate low frequency variants is replicated in the imputation meta-analysis (as seen for the first two variants listed in **Table 6.10**); at other variants, we do not observe replication. This could be used to further filter the list of candidate causal low frequency variants (after evaluating imputation quality at each site).

D) ***What is the role of rare variation, independent of the common GWAS signal?***

We identify only one Goldilocks (MAF > 0.01, OR > 2 or OR < 0.5, p-value < 0.01) variant within the HMGA2 GWAS locus, and 10 such variants across the broader 5-Mb interval (**Table 6.11**). Signal at none of these variants replicated in the imputation meta-analysis.

Gene-based association testing across the HMGA2 interval revealed a potentially interesting result at the HMGA2 gene (SKAT-O p=0.0021 for all non-synonymous variants with MAF<1%). This signal remained unchanged after conditioning on the GWAS tag SNP (p=0.0028), suggesting that it was not driven by association of an underlying common haplotype. The single missense variant driving this signal is listed in **Table 6.12**; it did not replicate in imputation meta-analysis, but its very low frequency (and thus uncertain imputation accuracy) still makes it a potential follow-up candidate.

Table 6.11: Potential Goldilocks' low frequency variants at the HMGA2 locus

Variant	MAF	Case Obs	Control Obs	OR	P-value	Conditional P-val	Protein-coding?	Imputation p-value (~50K samples)
12:66296807	1.1%	17	40	0.44	5.87E-03	7.97E-03	No	0.939
In the broader 5Mb interval:								
12:67792145	1.5%	57	23	2.56	1.65E-04	1.44E-04	No	0.234
12:67760420	1.5%	56	23	2.51	2.41E-04	2.13E-04	No	0.287
12:64262864	1.9%	32	69	0.47	6.29E-04	6.74E-04	No	0.630
12:65666150	1.8%	31	64	0.49	1.16E-03	1.46E-03	No	0.286
12:65668735	1.0%	16	39	0.39	1.80E-03	1.93E-03	No	0.377
12:65669291	1.0%	16	39	0.39	1.80E-03	1.93E-03	No	0.433
12:65666365	1.1%	17	39	0.42	2.95E-03	3.21E-03	No	0.490
12:65703224	1.0%	17	38	0.43	4.05E-03	4.07E-03	No	0.272
12:64750252	1.0%	17	37	0.43	4.83E-03	5.26E-03	No	0.368
12:65923839	1.3%	23	45	0.49	6.00E-03	8.57E-03	No	0.183

Table 6.12: Rare non-synonymous (MAF<1%) variants in HMGA2 gene driving burden signal

Variant	Case Obs	Control Obs	OR	P-value	Annotation; PolyPhen Category	Imputation p-value
12:66308896	8	28	0.30	0.003	Missense (p.K103E); unknown	0.47
12:66221812	1	0	-	-	Missense (p.P48R); probably damaging	-
Total	**9**	**28**				

No loss-of-function variants were identified in the HMGA2 gene.

E) *Summary*

Data at the HMGA2 locus indicate that the original GWAS signal is likely the main genetic association to T2D at this locus. This signal could be explained, however, by either common variant(s) or by as few as three low frequency variants (such as those listed in **Table 6.10**). Both of these genetic models need to be explored further; this is challenging at this locus, however, due to the very large number of both common and low frequency candidate causal variants. To experimentally differentiate between these hypotheses, for example, an attractive strategy may involve using genome engineering to knock in the entire risk or non-risk haplotype amplified from human cells, and then individually mutate only candidate low frequency variants. It may not be tractable to individually test the effect of each candidate common variant on the haplotype (as may be possible at a smaller locus with fewer common tag SNPs, such as chr9p21 or CCND2).

Locus #4: HNF1B; GWAS tag SNP **rs4430796** (17:36098040; GoT2D MAF= 0.43; OR = 1.19)

This tag SNP has 8 common SNPs in tight LD ($r^2 > 0.8$) and 23 common variants (21 SNPs, and 2 insertion/deletion variants) in modest LD ($r^2 > 0.5$). None of these variants are protein-coding. There is evidence for two independent common signals at this locus: one is tagged by the above listed SNP (and is localized to the first intron of HNF1B), while the other is located downstream of the HNF1B transcript within a hotspot of recombination. This second potentially novel signal (17:36022605; rs72830455; MAF = 0.11) shows strong association in the GoT2D panel (OR = 0.68; p=1.3e-05). Its replication is weak in imputation meta-analysis (p= 6.4e-04), but is still comparable to the signal at the original tag SNP (p-value after imputation at 17:36098040 = 1.2e-04). These two signals are independent of one another ($r^2 = 0.00$).

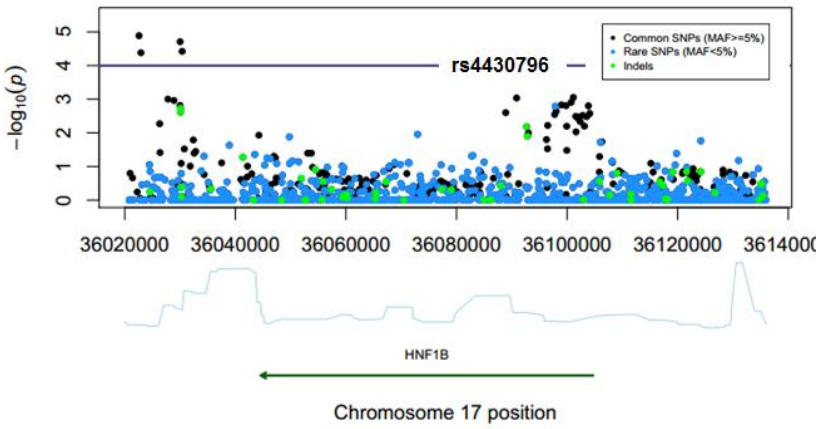


Figure 6.5: Association signal across the HNF1B GWAS locus

We asked whether rare or low frequency variants at this locus could explain the original GWAS signal (at 17:36098040). Similar to findings at TCFL2, we find that over 50 such variants would be required, making the most parsimonious model, by far, one implicating common causal variant(s). We did not perform this analysis for the second independent common variant signal.

Six 'Goldilocks' variants were identified across the 5-Mb interval; none of these fell within the GWAS interval defined by recombination hotspots, and none of these variants had signal which replicated in imputation meta-analysis. There was no significant burden of rare missense variants across HNF1B (SKAT-O p=0.295); no loss-of-function variants were identified.

Locus #5: IGF2BP2; GWAS tag SNP **rs4402960** (3:185511687; GoT2D MAF= 0.32; OR = 1.25)

Similar to the HMGA2 locus, this locus exhibits high haplotype complexity: the GWAS tag SNP has 76 common SNPs in tight LD (r^2 > 0.8) and 83 common variants (including 8 insertion/deletion variants) in modest LD (r^2 > 0.5). None of these variants are protein-coding; all are within an intron of IGF2BP2 or upstream of this transcript. After conditioning on the top common variant, no independent common variant associations are observed.

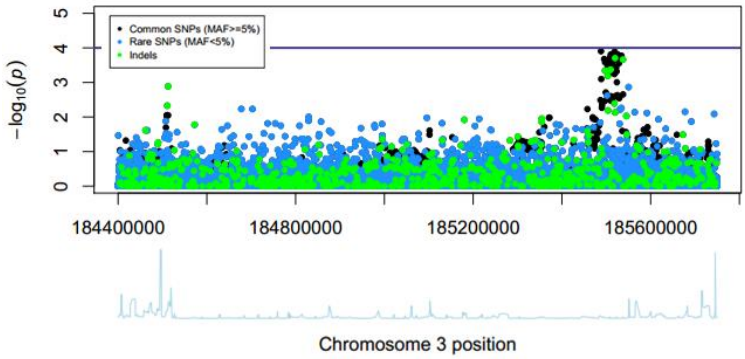


Figure 6.6: Association signal across the IGF2BP2 GWAS locus

Again similar to findings at HMGA2, we find that **as few as 6 low frequency variants** would be sufficient to explain the GWAS signal at the IGF2BP2 locus; in this case, the low frequency model actually has a *much greater likelihood* than a model with only the common haplotype (relative likelihood of common model = 0.0003, as assessed by AIC). An example of a set of low frequency variants that could produce synthetic association are listed in **Table 6.13**:

Table 6.13: Example low frequency variants sufficient to explain IGF2BP2 GWAS signal in GoT2D

Variant	MAF	MAC	Case Obs	Control Obs	OR	P-value in GoT2D
3:185550500	4.1%	219	133	86	1.57	1.36E-03
3:185530177	3.6%	192	111	81	1.44	0.02
3:185502055	2.1%	111	70	41	1.84	2.42E-03
3:185462028	1.6%	87	52	35	1.47	0.09
3:185272651	3.8%	200	111	89	1.22	0.17
3:185340766	1.8%	98	55	43	1.26	0.26

Four 'Goldilocks' variants were identified across the 5-Mb interval; none fell within the GWAS interval, but one (3:182657186; MAF=1.5%, OR=0.49 in GoT2D) had directionally consistent signal after imputation meta-analysis (p=0.007), potentially meriting follow-up. There was no burden of rare missense variants at IGF2BP2 (SKAT-O p=0.121); no LOF variants were identified in this gene.

Locus #6: PPARG; GWAS tag SNP **rs1801282** (3:12393125; GoT2D MAF= 0.15; OR = 0.82)

This GWAS tag SNP has 26 common SNPs in tight LD ($r^2 > 0.8$) and 32 common variants (including 2 insertion/deletion variants) in modest LD (r2 > 0.5). As is well-known at this locus[3], the original tag SNP is protein-coding (PPARG; p.P12A); the remaining variants in LD with the GWAS tag SNP are all non-coding and localize to an intron of PPARG. After conditioning on the top common variant, no independent signals are observed.

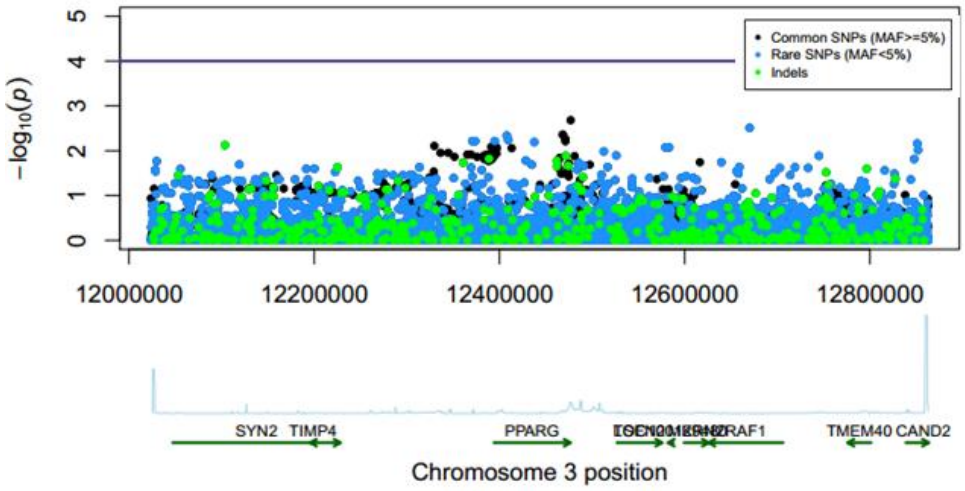


Figure 6.7: Association signal across the PPARG GWAS locus

We find that the protective effect observed at this GWAS haplotype (OR = 0.82) could be explained by as few as **6 low frequency variants**. An example of such a set of variants is shown in **Table 6.14** (none of these are protein-coding). In this case, the low frequency model has a greater likelihood of explaining the data relative to a model with only the common haplotype (~12 times more likely, as assessed by AIC). However, this should be interpreted with great caution given (a) that the common haplotype is tagged by a protein-coding missense variant with higher prior likelihood of altering function, (b) that the low frequency model is likely over-fit in this sample, and (c) that most of the candidate low frequency variants do not reproduce signal in imputation meta-analysis.

Table 6.14: Example low frequency variants sufficient to explain PPARG GWAS signal in GoT2D

Variant	MAF	MAC	Case Obs	Control Obs	OR	P-value in GoT2D	Imputation p-value (~50K samples)
3:12313747	1.8%	95	38	57	0.62	0.02	0.48
3:12086266	1.9%	102	44	58	0.72	0.11	not imputed
3:12757320	1.0%	51	19	32	0.62	0.10	0.27
3:12613046	0.5%	26	7	19	0.41	0.05	0.16
3:12381018	0.3%	16	3	13	0.24	0.03	0.51
3:12440205	2.9%	156	76	80	0.91	0.56	9.40E-03

Interestingly, we detect a nominally significant burden of rare non-synonymous variants at PPARG (SKAT-O p=0.045; p=0.051 after conditioning on the GWAS tag SNP; **Table 6.15**). This burden signal is driven by 8 extremely rare, case-private singletons, including one loss-of-function variant p.S249*. PPARG is the only gene across this GWAS locus that shows such a signal.

Table 6.15: Rare non-synonymous (MAF<1%) variants in PPARG gene driving burden signal

Variant	Case Obs	Control Obs	Annotation; PolyPhen Category
3:12421274	0	1	Missense (p.V52I); benign
3:12421355	1	0	Missense (p.E79K); benign
3:12434126	1	0	Missense (p.R165T); probably damaging
3:12434131	1	0	Missense (p.I167V); benign
3:12434133	1	0	Missense (p.I167M); probably damaging
3:12447507	1	0	Stop gained (p.S249*)
3:12458209	1	1	Missense (p.V276I); benign
3:12458335	1	0	Missense (p.V318M); probably damaging
3:12458516	1	0	Missense (p.R378K); probably damaging
3:12475403	1	0	Missense (p.P426L); probably damaging
Total	9	2	

Members of our laboratory are currently experimentally characterizing these and other non-synonymous variants in the PPARG gene (including the common, missense GWAS tag SNP) to understand both their potential mechanism of action as well as their association to a range of metabolic phenotypes. Across all ten T2D GWAS loci studied in this chapter, PPARG shows the strongest such burden of very rare disease-associated protein-coding variation.

Locus #7: CDKAL1; GWAS tag SNP **rs7756992** (6:20679709; GoT2D MAF= 0.31; OR = 1.12)

The CDKAL1 tag SNP has 12 common variants in tight LD (r^2 > 0.8) and 47 common variants (including 6 insertion/deletion variants) in modest LD (r2 > 0.5). None of these variants is protein-coding; all are non-coding and localize to an intron of CDKAL1.

In the GoT2D panel, association at the tag SNP is relatively weak (p=0.05). Visualization of association signal across the locus is somewhat deceptive: the top signal is actually a novel insertion/deletion mutation (located at 6:20730725; not seen in the 1000 Genomes Project) that is in only modest LD with the original tag SNP (r^2=0.21). Even after conditioning on the GWAS tag SNP, association at this variant remains (OR = 1.41, p= 1.49e-04). However, upon closer inspection this variant occurs in a poly-A track; it is an "AAAAGAAAG" insertion; there is a high chance this variant is an artifact. The top SNP signal (after excluding insertion-deletion events), however, still occurs at a site that is also in only weak LD with the original tag SNP: at 6:20741680 (r^2 = 0.21; association p=2.3e-05 in GoT2D panel). In the imputation meta-analysis, association at the tag SNP (p=1.8e-22) is significantly stronger than association at this second SNP (for which p=7.8e-08), but the second signal is nearly genome-wide significant. This suggests the possible presence of two independent common variant signals at the CDKAL1 locus.

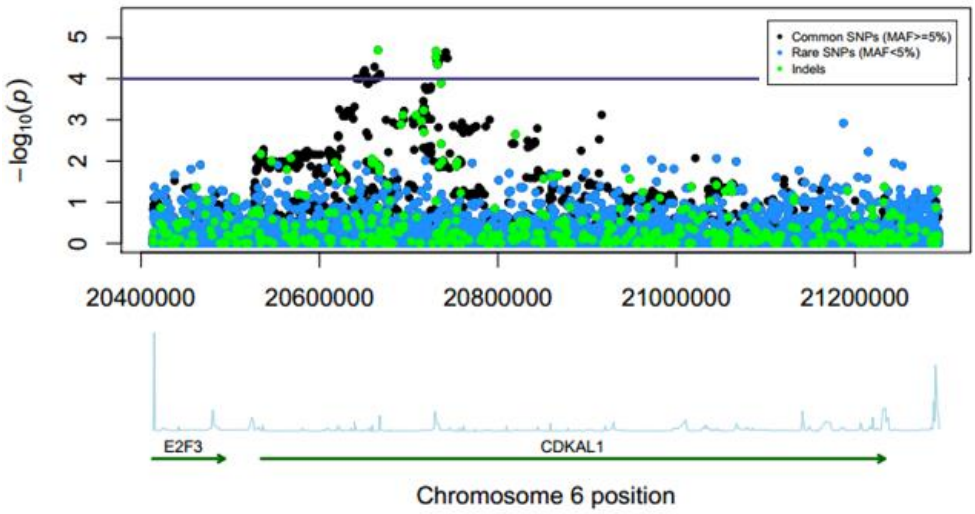


Figure 6.8: Association signal across the CDKAL1 GWAS locus

The effect at the original GWAS tag SNP was too weak to evaluate the possibility of synthetic associations in this dataset, but we did look for evidence for rare or low frequency effects independent of the GWAS signal. We identified 2 'Goldilocks' alleles in the 5-Mb interval; both were outside the GWAS locus, however, and neither replicated in imputation meta-analysis. A scan for a burden of rare missense variation across protein-coding genes identified a weak protective signal at the CDKAL1 gene (collapsed SCORE test p=0.051; p=0.066 after conditioning on the tag SNP). This signal is driven by the variants listed in Table 6.16 below; no loss-of-function variants were identified. Across a 5-Mb interval, CDKAL1 was the only gene showing nominal signal.

Table 6.16: Rare missense (MAF<1%) variants in CDKAL1 gene driving (weak) burden signal

Variant	Case Obs	Control Obs	Annotation; PolyPhen Category
6:20546697	16	26	Missense (p.R39Q); benign
6:20781441	1	1	Missense (p.D195Y); possibly damaging
6:20846371	0	1	Missense (p.I235T); possibly damaging
6:20955653	0	1	Missense (p.G249D); probably damaging
6:20955667	0	1	Missense (p.W254R); probably damaging
6:21000551	1	0	Missense (p.R335G); probably damaging
6:21201441	0	1	Missense (p.A495G); benign
6:21231263	0	1	Missense (p.Y578C); benign
Total	18	32	

Missense variants in CDKAL1 could be further evaluated in ongoing (larger) exome sequencing and genotyping studies to further test the (therapeutically attractive) hypothesis that missense variation in this gene is associated with protection against T2D.

Locus #8: JAZF1; GWAS tag SNP **rs849134** (7:28196222; GoT2D MAF= 0.48; no effect observed)

The tag SNP at the JAZF1 locus has 16 common variants in tight LD (r^2 > 0.8) and 22 common variants (including 4 insertion/deletion variants) in modest LD (r2 > 0.5). None of these variants is protein-coding; all are non-coding and localize to an intron of JAZF1. No signal is observed at the GWAS tag SNP in the GoT2D sequenced panel; thus we looked to the imputation meta-analysis to perform credible set analysis (as we did for chr9p21 in Chapter 5, and HMGA2 earlier in this chapter). This reveals that 12 of the 18 SNPs (insertion-deletion variants were not imputed) in modest LD with the tag SNP comprise a 95% credible set of candidate common causal variants (above black line in **Table 6.17**). These variants show essentially indistinguishable association signal, and either genetic studies in more samples or functional studies (see Chapter 7) will be required to identify individual variants with higher likelihood of being causal than others.

Table 6.17: Common SNPs in 95% credible set at JAZF1 locus, ranked by imputation p-value

Variant	MAF	r2 with GWAS tag	OR	P-value	Annotation	Imputation p-value (~50K samples)	Posterior probability
7:28200097	0.47	0.95	0.95	0.35	INTRONIC (JAZF1)	7.8E-07	0.128
7:28189411	0.49	0.94	0.96	0.47	INTRONIC (JAZF1)	8.7E-07	0.124
7:28198677	0.47	0.95	0.95	0.37	INTRONIC (JAZF1)	8.8E-07	0.112
7:28196413	0.47	0.95	0.95	0.36	INTRONIC (JAZF1)	1.1E-06	0.092
7:28209956	0.50	0.85	0.96	0.48	INTRONIC (JAZF1)	1.2E-06	0.085
7:28142088	0.50	0.79	0.94	0.27	INTRONIC (JAZF1)	1.4E-06	0.080
7:28185891	0.48	0.88	0.96	0.41	INTRONIC (JAZF1)	1.4E-06	0.075
7:28177338	0.49	0.84	0.95	0.32	INTRONIC (JAZF1)	1.8E-06	0.060
7:28196222	0.48	1.00	0.97	0.56	INTRONIC (JAZF1)	1.9E-06	0.060
7:28172732	0.49	0.84	0.95	0.37	INTRONIC (JAZF1)	1.9E-06	0.057
7:28192280	0.48	1.00	0.97	0.60	INTRONIC (JAZF1)	2.3E-06	0.049
7:28180556	0.49	0.93	0.97	0.63	INTRONIC (JAZF1)	3.2E-06	0.035
7:28194397	0.48	0.99	0.97	0.63	INTRONIC (JAZF1)	3.8E-06	0.028
7:28162674	0.48	0.79	0.96	0.46	INTRONIC (JAZF1)	9.0E-06	0.013
7:28179396	0.41	0.55	1.02	0.67	INTRONIC (JAZF1)	2.1E-04	0.001
7:28189549	0.42	0.66	1.02	0.75	INTRONIC (JAZF1)	2.5E-03	0.000
7:28183702	0.40	0.61	1.01	0.92	INTRONIC (JAZF1)	2.7E-03	0.000
7:28187806	0.42	0.65	1.01	0.82	INTRONIC (JAZF1)	2.7E-03	0.000

The top signal in the GoT2D sequenced panel actually occurs at a putative novel common (MAF=15%) site (7:28109834) which shows OR = 0.76 and p=3.4e-04. However, this variant shows little association after imputation meta-analysis (p=0.06), suggesting that it likely does not represent a true independent signal. We see no strong evidence of independent rare variants of large effect at this locus. There is no burden of T2D-associated protein-coding variation across the JAZF1 gene.

Locus #9: KCNJ11; GWAS tag SNP **rs5215** (11:17408630; GoT2D MAF= 0.44; OR = 1.14)

Like at PPARG, the tag SNP at the KCNJ11 locus is a missense variant in the KCNJ11 gene (p.V337I). The haplotype which this SNP tags is fascinating: there are 18 common variants in tight LD ($r^2 > 0.8$), of which *three* are missense variants! In addition to the tag SNP, 11:17408630 (p.K23E in KCNJ11; r2 to tag = 0.99) and 11:17408630 (p.A1369S in ABCC8; r2 to tag = 0.93) both show nearly identical association signal. It is particularly interesting that these variants span two different genes, especially since the KCNJ11 and ABCC8 proteins are known to interact with one another as part of a potassium channel complex in beta cells (rare penetrant mutations in either gene are known to cause neonatal diabetes). It is thus possible that a haplotype with three missense variants across these genes arose to common frequency due to epistasis (e.g., a synergistic biological effect of multiple mutations).

Signal at this locus was too weak to test for the possibility synthetic associations. We see no evidence for a burden of rare T2D-associated non-synonymous variation in either KCNJ11 (SKAT-O p=0.13) or ABCC8 (SKAT-O p=0.37). We detect one 'Goldilocks' variant within an intron of KCNJ11 (11:17392994; MAF = 1.1%; OR = 0.41 and p=0.002 in GoT2D), but this signal does not replicate in the imputation meta-analysis (p=0.08). At this locus, we conclude that the three common missense variants are likely the most attractive signals to follow-up experimentally.

Locus #10: KCNQ1; GWAS tag SNPs **rs231362** (11:2691471; GoT2D MAF = 0.47; OR = 0.93)

rs163184 (11:2847069; GoT2D MAF = 0.44; OR = 0.78)

The KCNQ1 locus (like chr9p21) is known to harbor two independent common variant association signals.[1,4] The first (rs231362) has 1 only common variant in modest LD with it (11:2692249; r2 = 0.58); thus this tags a haplotype with very few candidate common causal variants. In the GoT2D sequenced panel, we observe no association signal at either of these SNPs (p=0.19 and p=0.84); in the imputation meta-analysis, however, they both show comparable signal (p=1.0e-05 at 11:2692249 and p=1.8e-05 at 11:2691471).

At the second GWAS tag SNP (rs163184, which is completely independent from the first tag SNP and separated by a peak of recombination; r2 = 0.00), we observe strong association in the GoT2D samples (p=1.2e-05; **Figure 6.9**). This SNP has only 2 other common variants in tight LD with it (11:2838413 and 11:2844216) and 4 variants in modest LD (r2 > 0.5). The association of all these variants in imputation meta-analysis is shown in **Table 6.18**. None of these variants is protein-coding; all are non-coding and localize to an intron of KCNQ1.

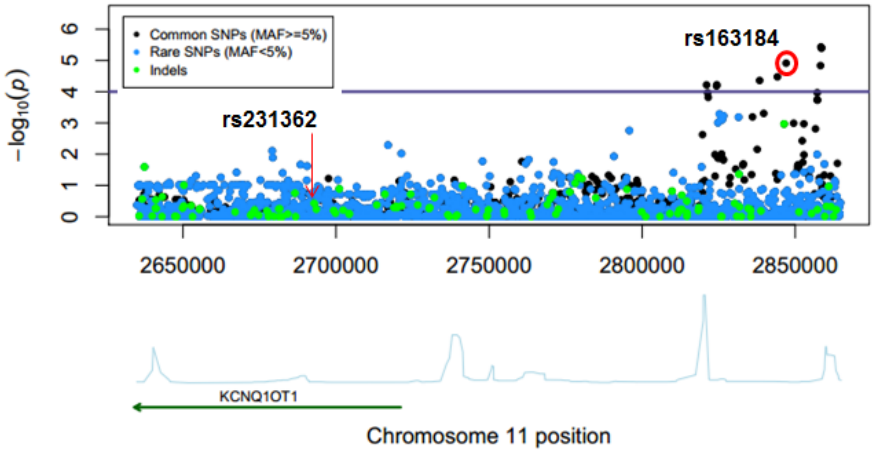


Figure 6.9: Association signal across the KCNQ1 GWAS locus

The meta-analysis results in particular raise the clear possibility that causal allele(s) at this locus could be either increasing or decreasing risk of T2D; the most associated variants after meta-analysis are actually SNPs for which the *major* allele is in modest LD with the tag SNP (and the minor allele shows association in the direction of increased T2D risk). It is never possible to infer with certainty, from only associated marker alleles, the *causal* direction of effect at any GWAS locus; the pattern of signals at KCNQ1 just highlights this challenge.

Table 6.18: Top common variants in LD with KCNQ1 GWAS tag SNP rs163184 (11:2847069)

			In GoT2D dataset			
Variant	MAF	r2 with 11:2847069	OR	P-value	Annotation	Imputation p-value (~50K samples)
11:2857194	46.5%	0.67	1.23	1.81E-04	INTRONIC (KCNQ1)	1.36E-09
11:2856658	44.5%	0.62	1.19	1.55E-03	INTRONIC (KCNQ1)	6.10E-09
11:2857297	49.9%	0.56	1.23	1.87E-04	INTRONIC (KCNQ1)	6.60E-09
11:2838413	42.7%	0.88	0.79	4.39E-05	INTRONIC (KCNQ1)	1.47E-08
11:2847069	44.5%	-	0.78	1.22E-05	INTRONIC (KCNQ1)	1.77E-08
11:2844216	46.8%	0.88	0.79	3.36E-05	INTRONIC (KCNQ1)	7.87E-07
11:2854514	42.5%	0.63	0.90	5.07E-02	INTRONIC (KCNQ1)	9.09E-04

Interestingly, however, none of this group of variants in LD with rs163184 is actually the *top* association signal observed in the GoT2D dataset. The top signal is a potentially *third* independent, relatively lower frequency signal which occurs at 11:2858440 (rs2237896; MAF = 5.5%; OR = 0.55, p=3.7e-06). This variant is present in the HapMap catalog, but may not have been on first generation GWAS arrays. After conditioning on this SNP, we see that association at the GWAS tag SNP rs163184 remains unchanged (and vice versa), suggesting that this is indeed an independent signal. This variant is in LD with only 5 other variants; these are listed in **Table 6.19** below. In imputation meta-analysis, one of these SNPs shows genome-wide significant association to T2D (p=2.2e-08), confirming that this appears to be a true third signal.

Table 6.19: SNPs tagging third independent, novel genome-wide significant signal at KCNQ1

			In GoT2D dataset			
Variant	MAF	r2 with 11:2858440	OR	P-value	Annotation	Imputation p-value (~50K samples)
11:2858546	5.8%	0.92	0.57	3.95E-06	INTRONIC (KCNQ1)	2.28E-08
11:2858440	5.5%	-	0.55	3.70E-06	INTRONIC (KCNQ1)	6.78E-08
11:2858636	5.5%	0.97	0.56	4.04E-06	INTRONIC (KCNQ1)	8.92E-08
11:2858295	5.1%	0.90	0.57	1.46E-05	INTRONIC (KCNQ1)	2.56E-07
11:2857233	7.2%	0.63	0.65	1.09E-04	INTRONIC (KCNQ1)	1.85E-05

We asked whether rare or low frequency variants might explain the signal at the second GWAS tag SNP (rs163184). We find that 15 variants segregating on the protective GWAS haplotype would be required to reduce the effect at the haplotype from 0.79 to 1.0, making a common causal variant the much more parsimonious genetic model. In this case, interpretation of this finding is made more challenging by the fact that some of the low frequency protective variants that are selected to explain the GWAS signal also partially tag the haplotype underlying the third independent signal. As a result, comparison of the explanatory power of the low frequency model to that of a model with only the common protective haplotype of rs163184 is confounded.

There are several (18) low frequency 'Goldilocks' alleles detected across the KCNQ1 locus, but in most cases these signals are at least partially explained by one of the two strong common variant signals at the locus (especially 11:2858440, which has a low MAF of 5.5%). Four of these variants show relatively strong association signal (p~1e-04 or less) in the imputation meta-analysis,

but their signal is much less than that observed at 11:2858440 (p=6.7e-08). In the GoT2D panel, these SNPs have the property that their association is reduced (from p=5e-06 to p=0.001, for example) but not entirely eliminated after conditioning on 11:2858440 (**Table 6.20**). All these four variants lie to the right of the GWAS interval (on the other side of a strong hotspot of recombination, up to 130kb away from 11:2858440), and exhibit only weak LD with 11:2858440 ($r^2 \sim 0.05$). To test the hypothesis that these low frequency variants might still be driving the association signal at 11:2858440, we tested association at 11:2858440 after conditioning on these four variants, and found that it was reduced but not eliminated (OR = 0.65, p=0.0016). These data collectively suggest that there might be additional, independent T2D association signals of low frequency at this locus. Some of these (Table 10.3) might be attractive candidates for follow-up genotyping in large samples.

Table 6.20: Potential 'Goldilocks' novel low frequency signals at the KCNQ1 locus

Variant	MAF	Case Obs	Control Obs	OR	P-value	OR, conditioned on 11:2858440	P-value, conditioned on 11:2858440	Protein-coding?	Imputation p-value (~50K samples)
11:2912406	2.2%	33	82	0.38	5.41E-06	0.49	1.78E-03	No	1.34E-04
11:2988846	2.8%	52	99	0.49	6.47E-05	0.57	1.62E-03	No	1.95E-05
11:2961671	1.4%	21	54	0.36	1.09E-04	0.45	2.62E-03	No	5.91E-05
11:2978772	2.7%	50	96	0.49	7.75E-05	0.57	1.91E-03	No	1.68E-05

Finally, we did not detect a significant burden of rare T2D-associated missense variation in the KCNQ1 gene (SKAT-O p=0.27).

References

1. Morris, A.P. *et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature Genetics* **44**, (2012).
2. Maller, J.B. *et al.* Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nature Genetics* (2012).doi:10.1038/ng.2435
3. Altshuler, D. *et al.* The common PPARγ Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nature Genetics* **26**, 76-80 (2000).
4. Voight, B.F. *et al.* Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nature Genetics* **42**, 579-89 (2010).

Chapter 7

Functional studies of regulatory elements across the chr9p21 and JAZF1 loci

The fine-mapping studies conducted in Chapters 5-6 narrow the space of genetic models plausible at any given T2D locus where genetic association is observed, but ultimately these hypotheses must be tested in biological systems in order to gain insight into mechanisms underlying the disease. One of the challenges we face in testing these hypotheses, however, is that in the vast majority of cases, the set of candidate causal variation across a GWAS locus lies entirely within non-protein-coding regions of the human genome. In this chapter, we present some (preliminary) experimental work, undertaken with Jessica Alston and Jason Wright, aimed at identifying non-coding regions with enhancer activity across the chr9p21 and JAZF1 loci. We reasoned that annotating function across an otherwise poorly understood genomic region would be a first step towards both prioritizing candidate causal variants within the locus and shedding light on some of the regulatory pathways underlying T2D pathogenesis.

Background: in vitro enhancer screening

Throughout this chapter, we use *in vitro* luciferase-based enhancer screens to identify fragments of the genome that may have regulatory function. Briefly, in this technique, a test fragment of DNA is cloned in a plasmid upstream of a luciferase reporter gene, and then transfected into cells of interest (see Methods below). The level of luciferase expression in cells transfected with the test fragment is compared to cells transfected with an 'empty' vector containing only a minimal promoter and the luciferase gene; test fragments that significantly increase or decrease expression of the luciferase reporter gene are identified as having potential *in vitro* enhancer or silencer activity. This technique has been broadly used in the literature to characterize non-protein-coding regions of the genome with putative regulatory function.

Of course, this screening method has many limitations in both its sensitivity and specificity. A biological enhancer's *in vivo* function may depend on many factors not recapitulated in an *in vitro* assay, such as the expression of particular genes and proteins (e.g. transcription factors), cellular state (e.g. dividing vs. senescent), and 3D chromatin structure (which may bring other parts of the genome in close proximity with a fragment to facilitate joint function). Moreover, the *in vitro* assay is typically performed in a cell line (into which DNA can be transfected) which likely does not behave exactly as human tissues in the body do. Thus, a fragment that does *not* show activity in this assay may very well still have endogenous *in vivo* activity. Specificity, too, may be lacking in this experiment: a fragment that shows enhancer activity when it is artificially cloned immediately upstream of a reporter gene may not actually have the ability to modulate expression of (perhaps farther away) genes in the endogenous human genome. Thus, all signals (both positive and negative) from such *in vitro* must be interpreted with great caution and must be further characterized (e.g. using newly available genome engineering techniques, for example, which enable manipulation of a test fragment in its endogenous genomic context; see **Appendix A4**).

Nonetheless, there are numerous examples in the literature of cases in which hits from an *in vitro* screen have enabled formulation of targeted biological hypotheses that could be highly informative on disease biology. In fact, at chr9p21, a study recently reported the discovery of such an enhancer fragment located within the region where common variants are associated to coronary artery disease (CAD).[1] This study identified an enhancer fragment containing a STAT1 binding site which appears to be disrupted by a CAD-associated variant. The authors also reported that activity of this regulatory unit was modulated by interferon-γ signaling, suggesting that this inflammatory pathway may play a role in CAD pathogenesis. This study did not report a systematic screen across the locus, and did not test all candidate causal variants for biological activity, but nonetheless presented a novel biological hypothesis that may merit further study. The extent to which such hypotheses will be validated as genome engineering approaches are more broadly applied remains

to be seen, but during the time this thesis was conducted, *in vitro* enhancer screens remained among the most informative windows into the function of non-protein-coding regions of the genome.

Biological background and motivation: chr9p21

The T2D association signal at chr9p21 is entirely contained within a relatively small non-protein coding region (as discussed in Chapter 5). It has been speculated that these variants may play a role in regulating the expression of the genes CDKN2A, which encodes $p16^{INK4a}$, and CDKN2B, which encodes $p14^{ARF}$, both of which are potent tumor suppressors. Evidence supporting this hypothesis comes from studies which demonstrate (a) that the targets of p16 are regulators of pancreatic beta-cell development[2,3], and (b) that p16 may play a role in beta-cell aging[4]. A causal variant which changes expression of these genes might cause dysregulation of cell cycle processes in beta cells, and might thereby cause the beta-cell burnout that is associated with the onset and progression of the human type 2 diabetes phenotype.

Alleles at the variants associated with CAD at chr9p21 (independent of the T2D-associated variants) have actually been shown to be correlated with expression of CDKN2A/B, lending support to the hypothesis that common variants may influence risk of common diseases by modulating expression of these genes. The T2D-associated alleles have not been correlated with CDKN2A/B expression in any studies conducted to date, but this absence of signal could be explained by the fact that few eQTL discovery studies have been performed in beta cells and other tissues relevant for T2D pathophysiology. While the chr9p21 region associated with CAD has been characterized via deletion in mouse[5] (and *in vitro* enhancer screening[1], as mentioned above), the T2D-associated region had not yet been experimentally interrogated for regulatory function when this work was initiated (by Jessica Alston).

Biological background and motivation: JAZF1

Like at chr9p21, common variants near JAZF1 have been associated to a host of human phenotypes including prostate cancer[6], human height, colon cancer, and lupus. All the candidate (common) causal variants identified in LD with the T2D-associated SNP (in Chapter 6) lie within the

first intron of the JAZF1 transcript. One attractive biological hypothesis is that variants within this intron are responsible for modulating expression of the JAZF1 gene. This hypothesis is strongly supported by the finding that T2D-associated common variants are also eQTLs for the JAZF1 transcript in adipose tissue; that is, alleles at these variants are associated with differences in the level of JAZF1 expression. We confirmed this finding by analyzing the effect of genotype at two T2D tag SNPs (rs1635853 and rs849140) on JAZF1 expression in publicly available expression data from 264 HapMap lymphoblastoid cell lines (**Figure 7.1**); these data suggest that the eQTL may be functional in many tissues (in addition to just adipose, where it was previously reported).

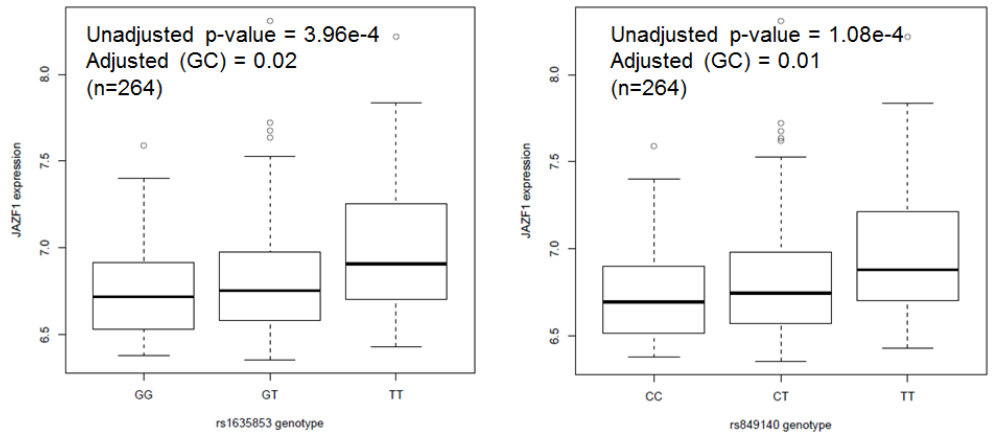


Figure 7.1: Effect of genotype at JAZF1 tag SNPs on JAZF1 expression in HapMap cell lines

Further evidence suggesting that JAZF1 may indeed be the gene target for T2D-associated regulatory variants comes from a recently described JAZF1 knock-out mouse, which has a metabolically dysregulated phenotype including postnatal growth retardation, reduced serum IGF1, excess fat accumulation, decreased muscle mass, and reduced insulin-stimulated whole body and muscle glucose uptake (unpublished; conference abstract). Thus, we might speculate that variants that increase risk of T2D do so by subtly increasing or decreasing expression levels of the JAZF1 gene in disease-relevant tissue types. The enhancer screens we conducted test this hypothesis.

Methods and experimental design

At both the chr9p21 and JAZF1 loci, we defined a panel of tiled test fragments, each of about 2-3kb in length, across the locus of interest. We were careful to ensure that GWAS tag SNPs were

included within these fragments, and also designed fragments to cover the boundaries of histone modification peaks across the region (annotated by the ENCODE Project). The test fragments designed for the chr9p21 T2D locus and the JAZF1 T2D locus are shown in **Figure 7.2** and **Figure 7.3** (the T2D GWAS tag SNPs highlighted at green dotted lines). As seen in both these figures, there is much reason to hypothesize that these regions may indeed contain regulatory functional elements. At chr9p21, **Figure 7.2** shows that the patterns of histone modification across multiple cell lines are consistent with the presence of several weak (yellow) and strong (orange) enhancer elements, at which many transcription factors have been shown to bind in ChIP-Seq experiments.

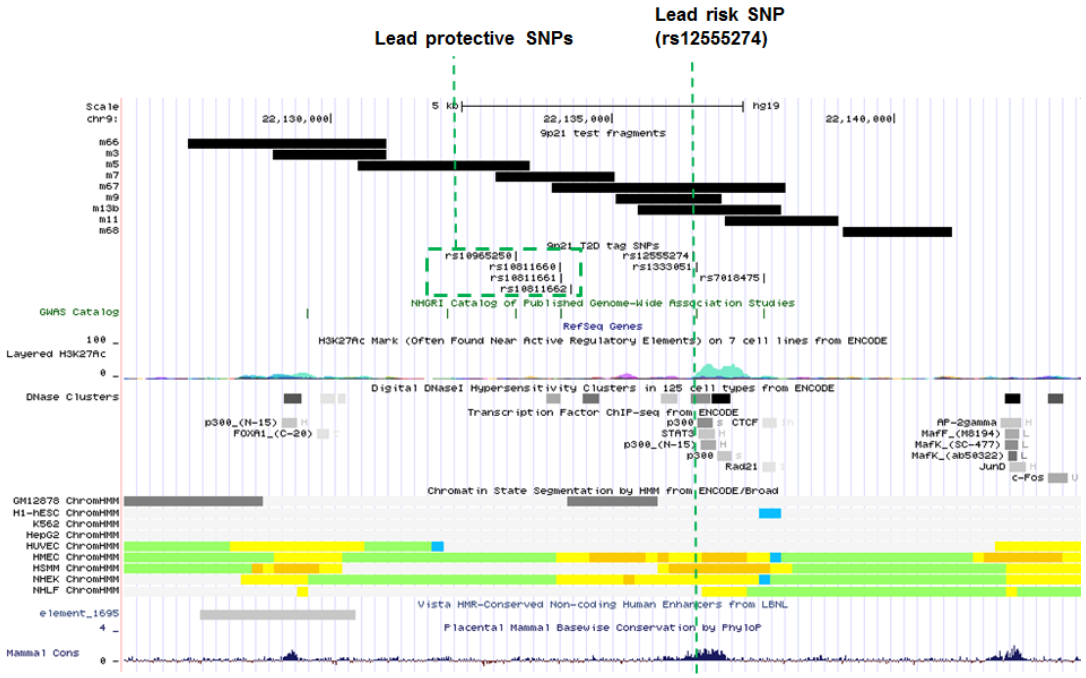


Figure 7.2: Test fragments designed across the T2D chr9p21 locus.
These are located in a non-protein-coding region of the genome >100kb away from the nearest gene.

In the first intron of JAZF1, we observe similar patterns: enhancer-associated histone modifications (e.g. H3K4Me3) are scattered across the (transcribed) intron across multiple cell types, and are also associated with the binding of numerous transcription factors (**Figure 7.3**). These observations further motivated the enhancer screening experiments described below.

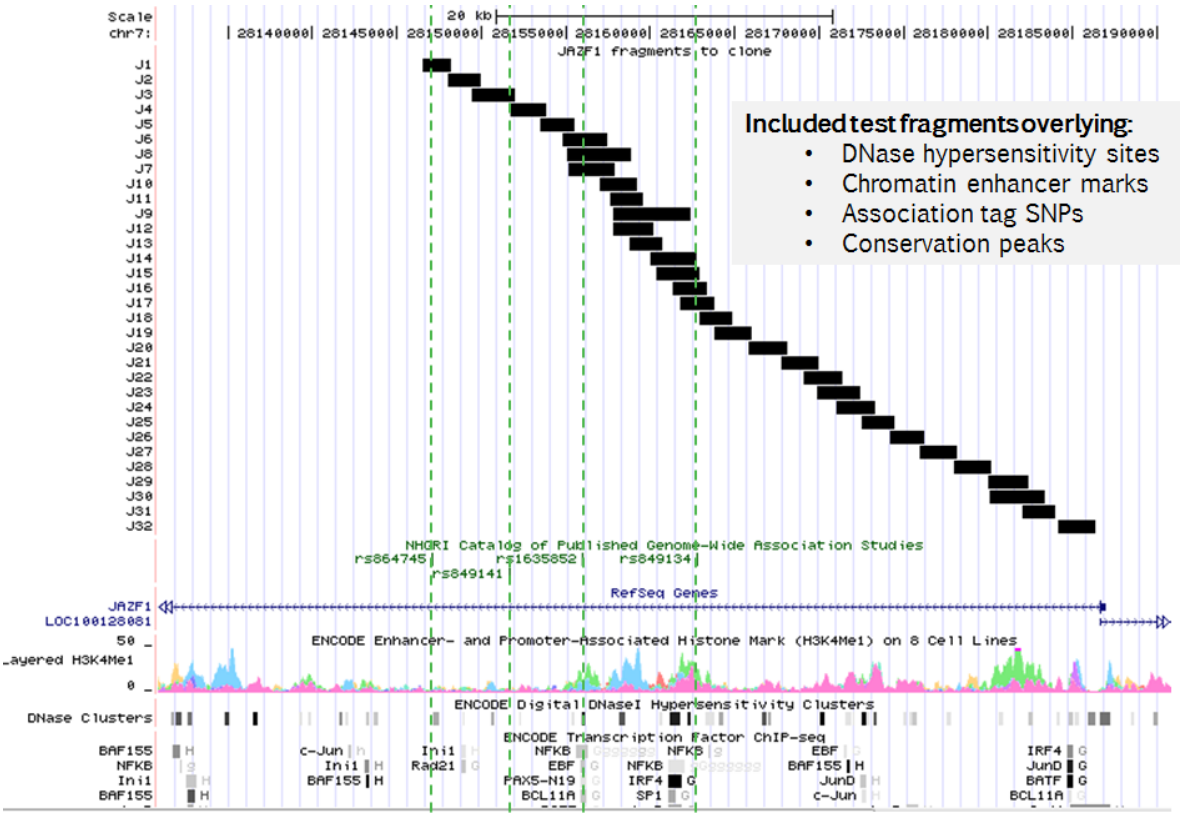


Figure 7.3: Test fragments designed across the T2D JAZF1 locus.
These are located within the first intron of the JAZF1 gene.

Each of the test fragments shown above was amplified from genomic DNA extracted from HapMap immortalized lymphoblastoid cell lines identified as being homozygous for the T2D-associated risk haplotype (different individual cell lines were used for chr9p21 and JAZF1 cloning). Each test fragment was then cloned upstream of a minimal promoter and firefly luciferase gene into the Promega pGL4.26 plasmid (**Figure 7.4**). Preps for all plasmids (at each locus) were produced on the same day, and concentrations were normalized before transfection to ensure maximal consistency across test fragments. In order to be able to interpret the enhancer activity of these test fragments, we recognized that positive and negative controls would be useful. These are difficult to define: there is not actually a vast literature on well-validated human enhancer elements, and the tissue-specificity of these elements makes it difficult to know which fragments likely have or lack pan-cell-type activity. To address this challenge, we leveraged a set of *in vivo* (transgenic mouse) enhancer screening data available as part of the

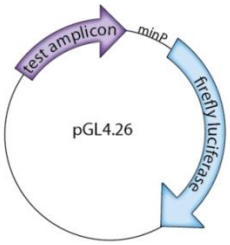


Figure 7.4: pGL4.26 plasmid used for enhancer screening

VISTA Enhancer Browser[7]. In this resource, hundreds of candidate human enhancer regions were cloned upstream of a LacZ reporter gene, injected into mouse oocytes, and then profiled for activity across a wide range of embryonic tissues. We used this dataset to identify fragments that showed constitutive activity across all embryonic tissues (one such fragment on chr9 is termed 'm14' below) and fragments showing no activity ('m15' below).

Each plasmid was transfected alongside a normalization control. This control plasmid contained a different luminescent reporter (renilla luciferase) downstream of a constitutively active promoter (in the pGL4.74 plasmid), and *no* test fragment. This co-transfection protocol (accomplished using Promega DualGlo reagents) enabled us to control for differences in transfection efficiency across master mixes, and differences in cell number across wells (of a 96-well plate). We optimized transfection conditions for a range of human cell lines (selected for a combination of ease of transfection as well as biological hypotheses about the cell types in which enhancer elements at chr9p21 and JAZF1 may function, respectively). The protocol we followed is roughly outlined below:

1. Plate cells in 96-well plate at ~10-20K cells per well.
2. Add pGL4.74 to DMEM (no additives) at a concentration of 200ng per ~87ul media.
3. Prepare each transfection master mix by mixing ~87ul DMEM+pGL4.74, 5ul test amplicon pGL4.26 plasmid (at 400ng/ul = 2ug total), and 8ul Fugene reagent.
4. Shake and let sit for 20 minutes.
5. Add 10ul of master mix to each well of cells, in which ~150ul media should be present.
6. Shake the transfected plate of cells, and incubate at 37C.
7. After ~36 hours, lyse cells and transfer to an opaque-bottom 96-well plate. Read out firefly luminescence, and after quenching read out renilla luminescence per well.
8. Record normalized (ratio of firefly to renilla) luminescence per well.

Results: chr9p21 studies

Much of the chr9p21 experiments were performed by Jessica Alston, who cultured and transfected a wide variety of cell types selected to represent the range of phenotypic associations reported across the broader chr9p21 locus: MCF7 cells for breast cancer, T98G and U87MG for glioma, A375 for melanoma, and HEPG2 for T2D. As part of this thesis, we developed protocols for

the culture and transfection of human primary pre-adipocytes, and performed the enhancer screen (in conjunction with Jason Wright) in rat pancreatic insulinoma cells (INS1E). The results of these screens across this broad panel of cell types are shown in **Figure 7.5** below.

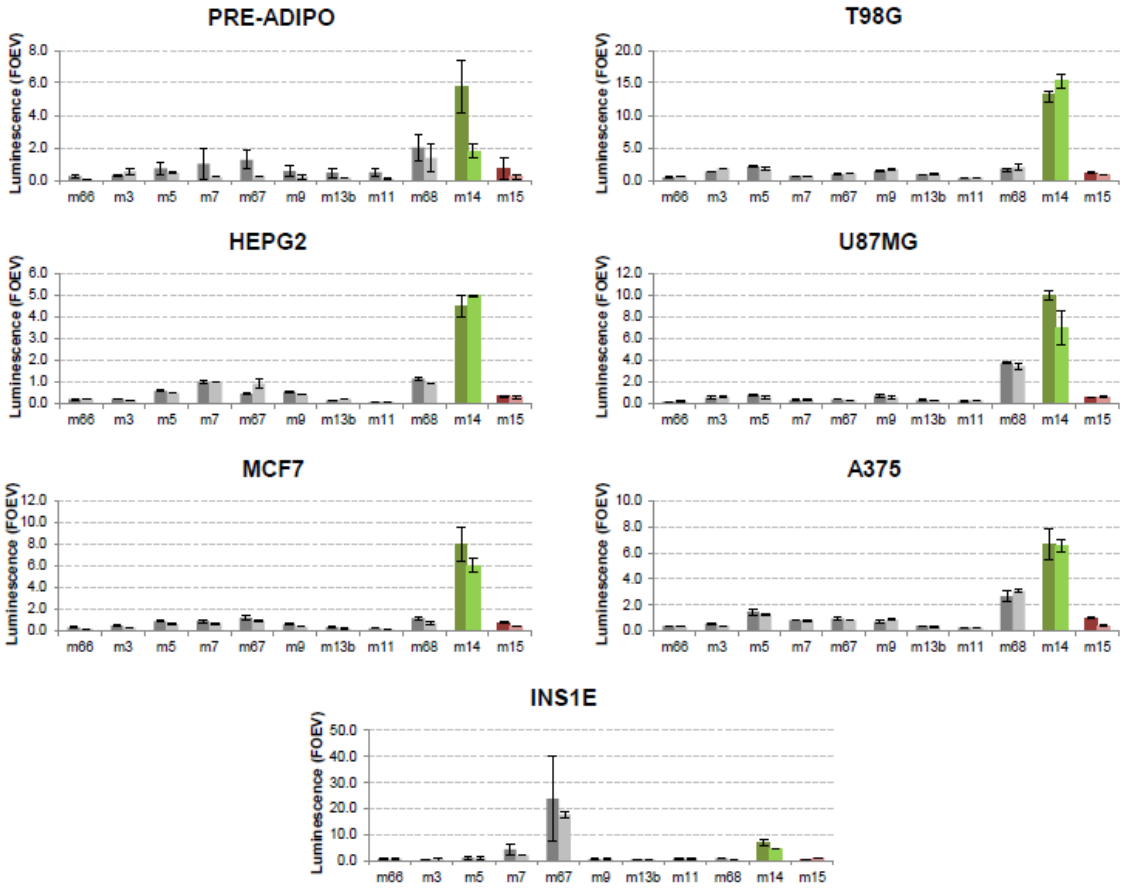


Figure 7.5: Enhancer screening across the chr9p21 T2D GWAS locus in a panel of cell lines identifies at a putative pancreas-specific regulatory element.

Two bars for each test amplicon represent independent transfection experiments performed on different plates of cells on different days. Standard deviations plotted for each bar represent range across wells of technical replicates.

The positive and negative controls (m14 and m15, respectively) show consistent activity across all cell types. No test fragment across the T2D locus shows pan-cell type activity, and in fact most fragments show no activity in any cell type. However, we identify a single fragment (m67) which shows strong and reproducible enhancer activity that is exclusive to the INS1E (pancreatic) cell line.

Partially overlapping amplicons or sub-fragments of the m67 amplicon (m7, m9, m13b, and m11) do not show activity, suggesting that the full fragment is required to exert enhancer function.

Table 7.1: Coordinates of chr9p21 enhancer amplicons, and candidate T2D variants within each fragment
SNPs tagging the risk haplotype are listed in orange; SNPs tagging the protective haplotype are in green.

Fragment	chr	start (hg19)	stop (hg19)	size (bp)	T2D-associated variants contained within fragment
m66	chr9	22127461	22130997	3536	
m3	chr9	22128965	22130988	2023	
m5	chr9	22130486	22133542	3056	rs10965250
m7	chr9	22132928	22135053	2125	rs10811661, rs10811662, rs10965250
m9	chr9	22135054	22136947	1893	rs12555274, rs1333051
m11	chr9	22136994	22139011	2017	rs7018475
m13b	chr9	22135453	22138001	2548	rs12555274, rs1333051, rs7018475
m67	chr9	22133932	22138083	4151	rs10811661 (on edge), rs10811662, rs12555274, rs1333051, rs7018475
m68	chr9	22139090	22141037	1947	

Interestingly, m67 happens to also contain several T2D candidate causal variants on both the risk and protective haplotypes (including rs12555274, which was identified in Chapter 5 as being >99% likely to drive the risk haplotype's association), raising the hypothesis that different alleles at these variants might confer different regulatory activity at this element (**Table 7.1**). Experiments to characterize the allele-specific activity of this fragment in INS1E cells – and also detect potential allele-specific differences in transcription factor binding – are currently being undertaken in our lab. We are also pursuing experiments to determine whether candidate transcription factors (such as HNF1B, which is predicted to bind at m67) are necessary for function of the element (e.g., by performing the screen in INS1E cells where HNF1B has been deleted). Given the observation of enhancer activity at the m67 fragment, members of our lab are also pursuing genomic deletion of this sub-fragment from the endogenous genome in human cells (in addition to deletion of the entire T2D GWAS locus) in an effort to connect regulatory elements to the expression of specific candidate genes (e.g. CDKN2A, CDKN2B, MTAP, or the non-coding transcript ANRIL).

Results: JAZF1 studies

At the JAZF1 T2D locus, we successfully cloned 19 out of the full panel of 32 test fragments designed in **Figure 7.3**. We were unable to clone the remainder due largely to PCR failures. Of the

19 fragments that were cloned, several contain candidate common causal SNPs at the JAZF1 locus (from **Table 6.8.1**; see **Table 7.2** below). In addition, during the time these experiments were designed, colleagues in Steve McCarroll's laboratory at the Broad Institute noticed a common structural variant at the JAZF1 locus (a 360bp deletion at chr7:28214300-28214664; hg19) in high LD with the protective allele at the T2D-associated GWAS tag SNP (r^2=0.91). A deletion of this size would be an interesting candidate causal variant, especially if it was overlying a region of putative regulatory function. We thus designed fragments J29 and J30 (**Figure 7.3**) specifically to ensure that they contained the deleted region (we cloned the risk haplotype of these fragments, which did not contain the deletion, with the hypothesis that the deletion might *reduce* the element's baseline regulatory activity). J29 and J30 were among the 19 amplicons that were successfully cloned.

We screened these 19 fragments for enhancer activity in only a single human cell line (293T; data shown in **Figure 7.6**). Most fragments show no activity, including J29 and J30. Four fragments, however, appear to have strong enhancer activity, increasing expression of the luciferase gene by more than 4-fold relative to the empty vector containing only a minimal promoter.

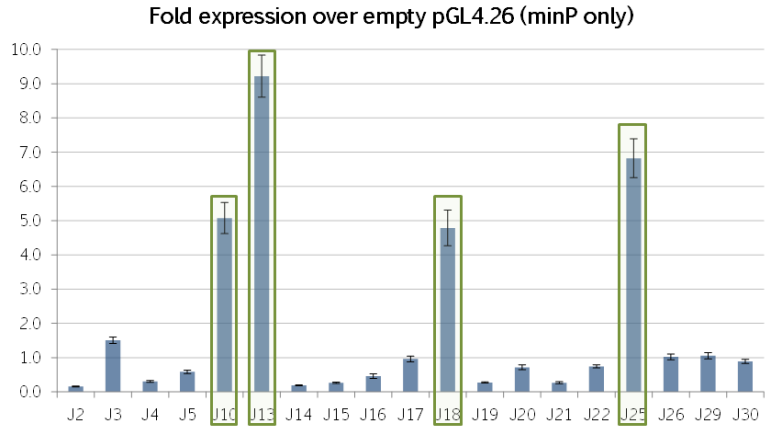


Figure 7.6: Enhancer screening across the JAZF1 locus with a subset of successfully cloned amplicons.
Error bars represent standard deviations across technical replicates.

These four fragments – J10, J13, J18, and J25 – contain several T2D candidate SNPs (which could now be tested for allelic effects by cloning these fragments from human cells containing the

protective rather than risk T2D haplotype), as well as a number of annotated transcription factor binding sites (**Figure 7.7**). These may provide interesting hypotheses to test in further studies.

Table 7.2 Coordinates of JAZF1 enhancer amplicons and candidate T2D variants within each fragment

Not yet successfully cloned

Fragment	chr	start (hg19)	stop (hg19)	size (bp)	T2D-associated variants
J1	chr7	28179975	28181675	1700	7:28180556
J2	chr7	28181475	28183475	2000	
J3	chr7	28182875	28185475	2600	
J4	chr7	28185175	28187275	2100	7:28185891
J5	chr7	28186975	28188975	2000	
J6	chr7	28188275	28190975	2700	7:28189411
J7	chr7	28188675	28191375	2700	7:28189411
J8	chr7	28188575	28192375	3800	7:28189411
J9	chr7	28191275	28195875	4600	
J10	chr7	28190475	28192675	2200	7:28192280
J11	chr7	28191075	28193075	2000	7:28192280
J12	chr7	28191275	28193675	2400	7:28192280
J13	chr7	28192275	28194175	1900	7:28192280 (on edge)
J14	chr7	28193475	28196175	2700	
J15	chr7	28193875	28196375	2500	
J16	chr7	28194775	28196875	2100	7:28196413, 7:28196222
J17	chr7	28195275	28197275	2000	7:28196413, 7:28196222
J18	chr7	28196375	28198375	2000	7:28196413
J19	chr7	28197275	28199475	2200	7:28198677
J20	chr7	28199275	28201575	2300	7:28200097
J21	chr7	28201275	28203475	2200	
J22	chr7	28202575	28204875	2300	
J23	chr7	28203375	28205875	2500	
J24	chr7	28204475	28206775	2300	
J25	chr7	28205975	28207975	2000	
J26	chr7	28207675	28209675	2000	
J27	chr7	28209475	28211675	2200	7:28209956
J28	chr7	28211475	28213675	2200	
J29	chr7	28213475	28215875	2400	Contains deletion region
J30	chr7	28213575	28216875	3300	Contains deletion region
J31	chr7	28215475	28217475	2000	
J32	chr7	28217675	28219875	2200	

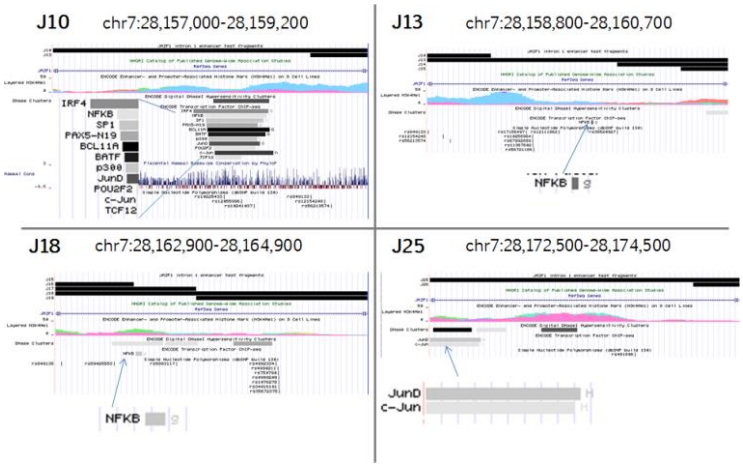


Figure 7.7: Predicted transcription factor binding and histone modifications across the four fragments showing strong enhancer activity at the JAZF1 T2D GWAS locus.

Very recently (May 2013), a study reported allele-specific differences in enhancer activity at the T2D JAZF1 GWAS tag SNP rs1635852 (chr7: 28189411).[8] The authors report a variety of allele-specific effects, including lower transcriptional activity, increased binding to protein complexes, and preferential binding to the pancreatic master regulator PDX1. Although the authors did not conduct a comprehensive screen for regulatory activity across the locus, and tested only 5 total SNPs in LD with the tag SNP (though there are 14 with $r^2 > 0.8$), these data are very compelling, and must be considered as any future experiments at the JAZF1 are planned. Unfortunately, this SNP falls within fragments J6, J7, and J8 in the enhancer screen designed here; these fragments are yet to be successfully cloned and so our data cannot support or reject this recently published hypothesis.

Discussion

In the set of experiments described above at chr9p21 and JAZF1, we have merely begun the process of annotating regulatory function across the T2D-associated locus. Our observations – of

islet-specific enhancer activity at a single fragment at chr9p21, and strong enhancer activity at four sub-fragments at JAZF1 – contribute to functional characterization of these regions, and suggest that these regions may indeed play a role in the regulation of gene expression. However, it is critical to recognize that these may just be the natural functions of these genomic regions, and the fragments we have identified may have nothing to do with T2D biology unless allele-specific effects are detected at SNPs that are associated with risk of T2D in human populations.

Nonetheless, we hope that the reagents developed during this work and the data presented in this chapter may help future investigators refine and test biological hypotheses at the chr9p21 and JAZF1 loci.

References

1. Harismendy, O. *et al.* 9p21 DNA variants associated with coronary artery disease impair interferon-γ signalling response. *Nature* **470**, 264-8 (2011).
2. Hino, S. *et al.* In vivo proliferation of differentiated pancreatic islet beta cells in transgenic mice expressing mutated cyclin-dependent kinase 4. *Diabetologia* **47**, 1819-30 (2004).
3. Rane, S.G. *et al.* Loss of Cdk4 expression causes insulin- deficient diabetes and Cdk4 activation results in β-islet cell hyperplasia. *Nature Genetics* **22**, 44-52 (1999).
4. Krishnamurthy, J. *et al.* p16INK4a induces an age-dependent decline in islet regenerative potential. *Nature* **443**, 453-7 (2006).
5. Visel, A. *et al.* Targeted deletion of the 9p21 non-coding coronary artery disease risk interval in mice. *Nature* **464**, 409-12 (2010).
6. Frayling, T.M., Colhoun, H. & Florez, J.C. A genetic link between type 2 diabetes and prostate cancer. *Diabetologia* **51**, 1757-60 (2008).
7. Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L.A. VISTA Enhancer Browser — a database of tissue-specific human enhancers. *Nucleic Acids Research* **35**, 88-92 (2007).
8. Fogarty, M.P., Panhuis, T.M., Vadlamudi, S., Buchkovich, M.L. & Mohlke, K.L. Allele-specific transcriptional activity at type 2 diabetes-associated single nucleotide polymorphisms in regions of pancreatic islet open chromatin at the JAZF1 locus. *Diabetes* **62**, 1756-62 (2013).

Chapter 8

Classes of variation genome-wide enriched for association to T2D

In Chapter 7, we studied two individual loci at which genetic variants associated with risk of T2D are hypothesized to exert their effect by changing the function of non-protein coding regulatory elements. This hypothesis is supported not only by the preliminary data we presented, but also by the observation that these loci are densely annotated with biochemical marks ('epigenetic modifications') suggestive of regulatory activity. Ultimately, more experimental studies will be required to test whether T2D-associated variants do in fact modulate enhancer activity at these individual loci, but one parallel way to gain insight into the *global* regulatory mechanisms underlying T2D is to ask the reverse question: genome-wide, are variants that lie within putative regulatory regions enriched for association to T2D?

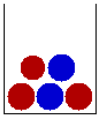
The ability to ask this question has only recently been enabled by the vast datasets generated by public consortia including the ENCODE and Epigenome Roadmap Projects[1]. Over the past few years, these projects have conducted thousands of experiments to profile histone modifications, DNase I hypersensitivity, transcription factor binding, gene expression, and more across a large number of human cell types. Patterns of co-occurrence and spatial relationships within these data have then been used to build genome-wide 'chromatin state' maps, or maps of regulatory state, for each cell type profiled.[2] A host of recent publications have leveraged these data to demonstrate that common variants associated to human phenotypes in GWAS are more likely (than expected by chance) to lie within regulatory elements.[3–6] For type 2 diabetes, for example, common SNPs at GWAS loci were shown to preferentially fall within ChIP-Seq peaks of the histone modification H3K4Me3 that occur in pancreatic islets and adult liver.[5]

These studies were restricted, however, to study of only the most strongly associated (genome-wide significant) loci for each trait, and in some cases they considered only incomplete catalogs of common variation (e.g. only variants present in the HapMap Phase 2 and 3 catalog). In

this chapter, we analyzed the whole-genome sequencing dataset described in Chapters 5-6 (which tests a complete catalog of *genome-wide* common and low frequency variation for association to T2D) to ask whether any regulatory classes (e.g. enhancers in any particular cell type) show evidence for enrichment for association to T2D. We present results for only common (MAF > 5%) variants below.

Method for evaluating enrichment

Another way of stating the above 'enrichment' question is simply: do variants falling in regulatory regions have association p-values that are skewed to be less than p-values for variants outside regulatory regions? This 'skew' can then be quantified and evaluated for statistical significance using a number of methods, including a hyper-geometric test performed at a given association p-value threshold (e.g. p=0.01). This test is described in **Figure 8.1** below. Intuitively, this test asks whether the number of variants in regulatory regions that have a T2D association p-value <0.01 is unusual, given the number of total variants that exist with p<0.01 and the number of variants in regulatory regions across the genome.



If I am drawing **n** balls from an urn with **N** balls, what is the chance I will draw $\mathbf{k_A}$ blue balls, given that there are $\mathbf{K_A}$ total blue balls in the urn?

$$hypergeometric\ p-value = \sum_{i=k_A}^{K_A} \frac{\binom{K_A}{i} * \binom{N-K_A}{n-i}}{\binom{N}{n}}$$

$n = total\ number\ of\ variants\ in\ regulatory\ regions$
$N = number\ of\ total\ variants\ genome-wide$
$K_A = total\ number\ of\ variants\ genome-wide\ with\ p < 0.01$
$k_A = number\ of\ regulatory\ variants\ with\ p < 0.01$

Figure 8.1: Hyper-geometric test for enrichment at T2D association p-value threshold of 0.01

We performed this hyper-geometric test at a variety of p-value thresholds, for common variants lying in regions annotated as enhancers, DNase I hypersensitivity sites, and FAIRE-Seq peaks (another open chromatin mark suggestive of regulatory activity) in a panel of 15 different human cell types (the cell lines GM12878, HepG2, NHEK, HUVEC, K562, HeLa, and HI-hESC, as well as primary cell types including pancreas, islet, adipose, smooth muscle, kidney, skeletal muscle, and colon).

Results

The results of these analyses are shown in **Figure 8.2-3**. Briefly, we detect strong enrichment signals in regulatory regions annotated as enhancers across a number of cell types, including pancreas, GM12878 (blood), HepG2 (a liver cell line), adipose tissue, smooth muscle, and kidney. We also detect (weaker) enrichment for T2D association across common variants in islet FAIRE-Seq peaks. It is interesting that this set of cell types contains numerous tissue types already thought to play a role in T2D biology (e.g., islets, liver, and adipose).

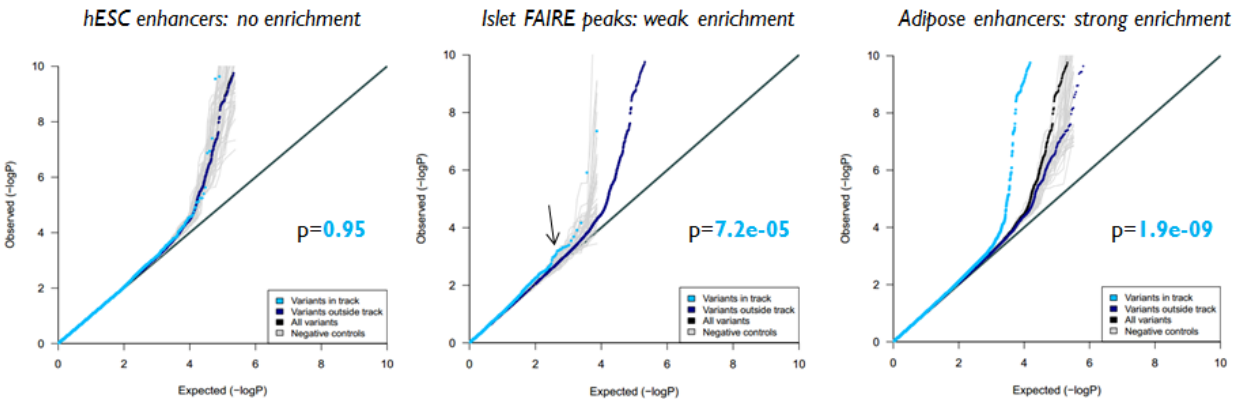


Figure 8.2: Examples of regulatory annotations that are not enriched, weakly enriched, or strongly enriched for common variant association to T2D.

Quantile-quantile plots show all common variants intersecting the regulatory annotation (e.g. common variants falling within regions annotated as enhancers in hESCs) in light blue, variants not intersecting the track in dark blue, and all common variants genome-wide in black. (Often the dark blue and black are indistinguishable if the regulatory annotation covers only a very small part of the genome). Light grey lines (there are 50 in each plot) represent randomly generated negative control regulatory annotations, generated to match the test annotation in total number and frequency distribution of overlapping variants. Hyper-geometric test is performed at an association p-value threshold of 0.01 as shown in **Figure 8.1**. As seen visually, significant enrichment is detected when the light blue points diverge from both the dark blue points, as well as from the negative controls (light grey).

The multitude of cell types that show enrichment was initially surprising (and in contrast to the tissue-specific enrichment results previously described[5] at only the top T2D GWAS loci). However, regulatory regions are often active across a wide range of cell types, and thus the annotations may overlap a great deal; we have not yet characterized the degree of this overlap. This does not necessarily mean that a large number of tissue types must be implicated in T2D biology (though this is certainly possible); even if T2D variants lie in regions with broad activity,

allele-specific effects on function could only exist within certain cell types (due to the presence of particular factors expressed in these cells, for example).

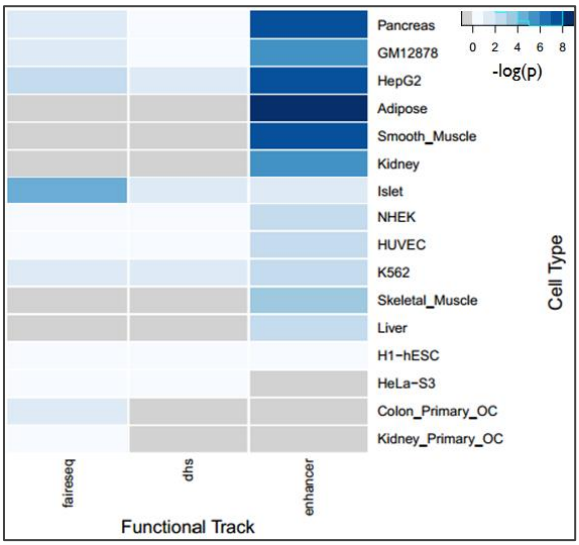


Figure 8.3: Results of testing for enrichment for association to T2D at common variants across non-coding regulatory regions annotated in a panel of human cell types.

Heat map shows $-\log_{10}$(hyper-geometric p-value) computed at an association p-value threshold of 0.01. Grey indicates no regulatory annotation was available for this cell type. Dark blue indicates more significant enrichment.

Given that the top T2D GWAS regions have been previously shown to be enriched for variants in regulatory regions, we wondered to what extent the genome-wide enrichments we observed were driven by already known GWAS signals vs. potentially novel regions of association. To assess this, removed all variants within 200kb of any T2D GWAS tag SNP or within 50kb of the protein-coding genes nearest GWAS tag SNPs. We then re-computed hyper-geometric p-values and re-visualized the skew in T2D association p-values (as was done in **Figure 8.2**). We find that for some regulatory annotations (e.g. the adipose enhancers), the enrichment is largely driven by GWAS loci (after removing GWAS loci, hyper-geometric p=0.22 as compared to 1.9e-09 before). However, for other annotations, such as HepG2 enhancers, significant enrichment remains even after removal of GWAS loci (hyper-geometric p=6.4e-06 after removal; see **Figure 8.4**). This suggests that a broad set of common variants across HepG2 enhancers genome-wide might each be contributing weak effects to the inherited risk of T2D.

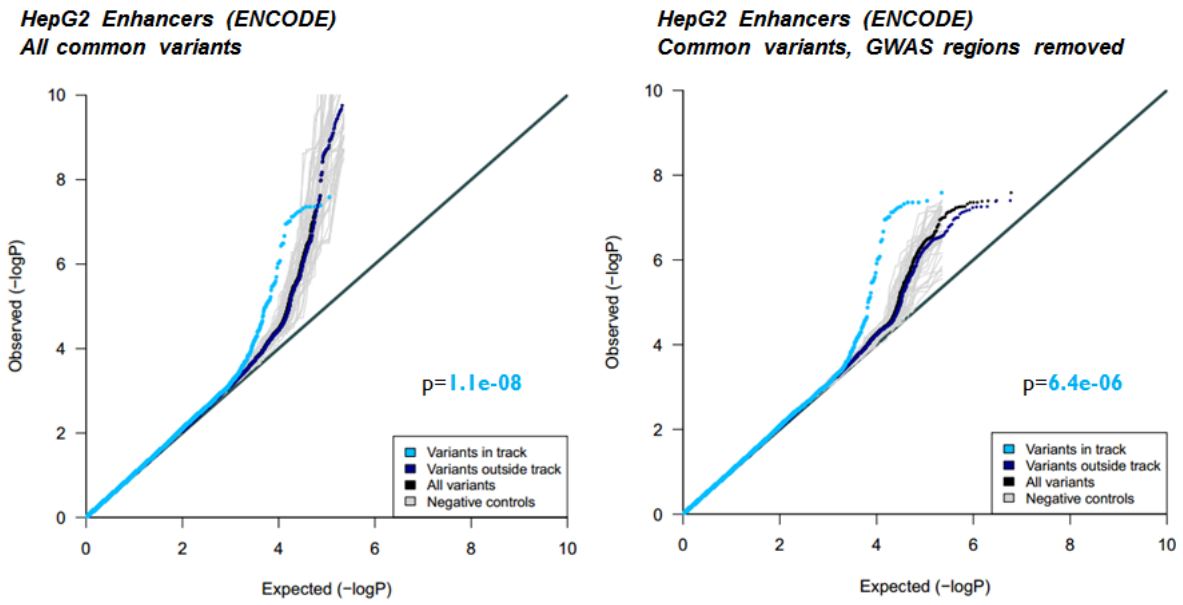


Figure 8.4: Example of enrichment (at HepG2 enhancers) not driven entirely by T2D GWAS loci

Figure at left is shown for all common variants genome-wide (as in **Figure 8.2**). At right, common variants in T2D GWAS regions have been removed.

These findings, if true, have several implications. First, the patterns of enrichment seen across multiple regulatory annotations would be consistent with genetic models in which a very large number of causal variants are spread across a large number of regulatory regions. Because these variants are likely to alter gene expression rather than protein function (and that too in subtle ways, given the redundancy of regulatory elements), we might expect that the effect sizes of these variants are weak. Second, the discovery of regulatory annotation classes that are clearly enriched for T2D association may improve power to discover novel disease associations in genetic studies; the prior probability that variants detected within such regulatory elements are associated to T2D would be higher, and this may facilitate more informed and targeted genetic follow-up (e.g. of a few variants, in larger sample sizes). Finally, the identity of the cell types showing enrichment for T2D association may be useful for functional investigators studying particular genes and pathways. Regions containing T2D-associated variants could be scanned for regulatory elements in the tissue types identified here; the presence of such elements may not only prioritize candidate causal variants, but also point to cell types in which these variants might be functionally interrogated.

Much further analysis is required, however, to further test and confidently interpret these findings. For example, we have not yet accounted for linkage disequilibrium patterns between common variants. If the haplotypes at disease-associated variants in regulatory regions are longer than average haplotypes, for example, this could explain a portion of the enrichment observed (many SNPs highly correlated with each other could be driving the top signals). Such a finding would not negate the enrichment altogether, but may impact its interpretation (with respect to questions about the number of unique enhancers contributing to risk of T2D, for example).

Additional tests of statistical significance (e.g., Mann Whitney rank-sum test, Kolmogorov-Smirnov test) should be performed to further characterize the nature and robustness of the enrichments observed. In particular, permutation-based approaches may also be informative: a distribution of negative controls (such as those depicted as grey lines in **Figures 8.2**) might be quantitatively compared to the observed set of p-values across a particular functional annotation category. These negative controls could be randomly selected genome-wide SNPs matched on various properties (e.g. frequency, distance from genes, haplotype length), or they could be generated by locally shifting the regulatory annotations (e.g. to preserve locus properties, but evaluate whether common SNPs specific to those elements are actually producing excess association signal). Future studies should also evaluate enrichment across low frequency and rare variants, as well as in regulatory annotations across a broader panel of cell types and tissues.

References

1. Frazer, K. a Decoding the human genome. *Genome Research* **22**, 1599-601 (2012).
2. Ernst, J. & Kellis, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature Computational Biology* **28**, 817-25 (2010).
3. Maurano, M.T. *et al*. Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science* **337**, 1190-1195 (2012).
4. Schaub, M. a, Boyle, A.P., Kundaje, A., Batzoglou, S. & Snyder, M. Linking disease associations with regulatory information in the human genome. *Genome Research* **22**, 1748-59 (2012).
5. Trynka, G. *et al.* Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nature Genetics* **45**, 124-30 (2013).
6. Gerstein, M.B. *et al.* Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**, 91-100 (2012).

Future directions

In this thesis, we have described various methods aimed at integrating diverse datasets to gain insight into the underlying architecture of disease type 2 diabetes (T2D). These included a population genetic simulation framework, haplotype-based fine-mapping techniques, and approaches for interrogating experimental data at non-protein-coding disease-associated regions of the genome. The work begun here is far from complete, and a number of exciting avenues of future research follow from the results presented in this thesis:

(1) Extensions of the population genetic simulation framework.

The simulations described in Chapters 1-3 could be improved to model more features of human genetic data in a number of ways: diverse locus structures could be simulated, non-protein coding regions could be simulated around genes (with different distributions of purifying selection), and multi-ethnic population histories could be simulated to evaluate genetic studies conducted in non-European populations. Such extensions can be made readily by training a few novel simulation parameters and using the methods already presented here.

(2) Incorporation of data from recently completed T2D genetic studies into the framework.

Chapter 3 compared the results of simulated genetic studies – linkage and GWAS – to those generated empirically for T2D, and generated predictions for the results of ongoing sequencing and large-scale genotyping studies under different disease models. In particular, we predicted that the results of such studies would diverge significantly under models that currently appear to be consistent with the results of GWAS. As this thesis is being written, analysis of these sequencing studies is ongoing and very soon these data could be rigorously compared to simulations under disease models that could not yet be excluded at the end of Chapter 3. These data may place much tighter bounds on the genetic architecture of T2D.

(3) Follow-up genotyping and experimental characterization of variants identified in fine-mapping studies.

In order to translate the genetic results presented in Chapters 5-8 into biological mechanisms of T2D, a great deal of follow-up is required. We hope that the results presented here constrain, to some extent, the full space of plausible hypotheses that must be tested. At some loci, for example, we describe a relatively small set of candidate causal mutations that might each be engineered in human cells (perhaps in tissue types that showed enrichment for association to T2D in Chapter 8). In other cases, we identify a handful of putative low frequency variants that might be prioritized for genotyping in a follow-up study. Finally, at chr9p21 and JAZF1, the regulatory elements we identified could be tested for allele-specific effects and potentially connected to biological pathways.

(4) Application of these methods to the study of other complex traits.

All the methods described here are applicable to the study of many complex traits. Linkage and GWAS have been conducted for a host of common diseases, and genotyping and sequencing studies are now underway. For every trait, integrating data from these diverse studies (alongside epidemiological data on the prevalence and heritability of the trait) presents a daunting challenge; specifying precise hypotheses that are grounded in population genetic principles offers one route to insight about the underlying genetic architecture. If the framework we describe in this thesis were applied to a range of complex traits for which different patterns have been observed in GWAS (e.g. autism, inflammatory bowel disease, AMD), it would be fascinating to compare and contrast the constraints placed by empirical data on each trait's genetic architecture.

These are just some of the most immediate follow-on research projects apparent to us at the present time; readers of this thesis will hopefully be inspired to forge others! Even though the analytical and experimental methods available to investigators are likely to change in coming years, we hope that the framework for hypothesis testing presented here may retain enduring value for future research in the field of human genetics.

Appendices

Appendix A1

ForSim configuration file containing best-fit evolutionary parameters

Below is the configuration file containing the mutation and recombination rates, demographic history parameters, and selection coefficient distribution which produced the best match to empirical sequencing data (as described in Chapter 1). This file can be used with the forward simulation software tool ForSim[1] to generate empirically calibrated human genetic sequence variation in large sample sizes.

References

1. Lambert, B.W., Terwilliger, J.D. & Weiss, K.M. ForSim: a tool for exploring the genetic architecture of complex traits with controlled truth. *Bioinformatics* **24**, 1821-2 (2008).

Configuration file

```
global begin
fertility poisson 2.0

2.0 megabases per centiMorgan male
2.0 megabases per centiMorgan female
mutation rate 2.0 e -8.0 female
mutation rate 2.0 e -8.0 male

generations 50371
event 50000 setCarryingCapacity Population 2000
event 50001 setCarryingCapacity Population 2026
event 50002 setCarryingCapacity Population 2052
event 50003 setCarryingCapacity Population 2078
event 50004 setCarryingCapacity Population 2105
event 50005 setCarryingCapacity Population 2132
event 50006 setCarryingCapacity Population 2159
event 50007 setCarryingCapacity Population 2187
event 50008 setCarryingCapacity Population 2215
event 50009 setCarryingCapacity Population 2244
event 50010 setCarryingCapacity Population 2273
event 50011 setCarryingCapacity Population 2302
event 50012 setCarryingCapacity Population 2332
event 50013 setCarryingCapacity Population 2362
event 50014 setCarryingCapacity Population 2392
event 50015 setCarryingCapacity Population 2423
event 50016 setCarryingCapacity Population 2454
event 50017 setCarryingCapacity Population 2486
event 50018 setCarryingCapacity Population 2518
event 50019 setCarryingCapacity Population 2550
```

```
event 50020 setCarryingCapacity Population 2583
event 50021 setCarryingCapacity Population 2616
event 50022 setCarryingCapacity Population 2650
event 50023 setCarryingCapacity Population 2684
event 50024 setCarryingCapacity Population 2719
event 50025 setCarryingCapacity Population 2754
event 50026 setCarryingCapacity Population 2789
event 50027 setCarryingCapacity Population 2825
event 50028 setCarryingCapacity Population 2861
event 50029 setCarryingCapacity Population 2898
event 50030 setCarryingCapacity Population 2935
event 50031 setCarryingCapacity Population 2973
event 50032 setCarryingCapacity Population 3011
event 50033 setCarryingCapacity Population 3050
event 50034 setCarryingCapacity Population 3089
event 50035 setCarryingCapacity Population 3129
event 50036 setCarryingCapacity Population 3169
event 50037 setCarryingCapacity Population 3210
event 50038 setCarryingCapacity Population 3251
event 50039 setCarryingCapacity Population 3293
event 50040 setCarryingCapacity Population 3335
event 50041 setCarryingCapacity Population 3378
event 50042 setCarryingCapacity Population 3422
event 50043 setCarryingCapacity Population 3466
event 50044 setCarryingCapacity Population 3511
event 50045 setCarryingCapacity Population 3556
event 50046 setCarryingCapacity Population 3602
event 50047 setCarryingCapacity Population 3648
event 50048 setCarryingCapacity Population 3695
event 50049 setCarryingCapacity Population 3743
event 50050 setCarryingCapacity Population 3791
event 50051 setCarryingCapacity Population 3840
event 50052 setCarryingCapacity Population 3889
event 50053 setCarryingCapacity Population 3939
event 50054 setCarryingCapacity Population 3990
event 50055 setCarryingCapacity Population 4041
event 50056 setCarryingCapacity Population 4093
event 50057 setCarryingCapacity Population 4146
event 50058 setCarryingCapacity Population 4199
event 50059 setCarryingCapacity Population 4253
event 50060 setCarryingCapacity Population 4308
event 50061 setCarryingCapacity Population 4363
event 50062 setCarryingCapacity Population 4419
event 50063 setCarryingCapacity Population 4476
event 50064 setCarryingCapacity Population 4534
event 50065 setCarryingCapacity Population 4592
event 50066 setCarryingCapacity Population 4651
event 50067 setCarryingCapacity Population 4711
event 50068 setCarryingCapacity Population 4772
event 50069 setCarryingCapacity Population 4833
event 50070 setCarryingCapacity Population 4895
event 50071 setCarryingCapacity Population 4958
event 50072 setCarryingCapacity Population 5022
event 50073 setCarryingCapacity Population 5087
event 50074 setCarryingCapacity Population 5153
event 50075 setCarryingCapacity Population 5219
event 50076 setCarryingCapacity Population 5286
```

```
event 50077 setCarryingCapacity Population 5354
event 50078 setCarryingCapacity Population 5423
event 50079 setCarryingCapacity Population 5493
event 50080 setCarryingCapacity Population 5564
event 50081 setCarryingCapacity Population 5636
event 50082 setCarryingCapacity Population 5709
event 50083 setCarryingCapacity Population 5783
event 50084 setCarryingCapacity Population 5857
event 50085 setCarryingCapacity Population 5932
event 50086 setCarryingCapacity Population 6008
event 50087 setCarryingCapacity Population 6085
event 50088 setCarryingCapacity Population 6163
event 50089 setCarryingCapacity Population 6242
event 50090 setCarryingCapacity Population 6322
event 50091 setCarryingCapacity Population 6403
event 50092 setCarryingCapacity Population 6485
event 50093 setCarryingCapacity Population 6569
event 50094 setCarryingCapacity Population 6654
event 50095 setCarryingCapacity Population 6740
event 50096 setCarryingCapacity Population 6827
event 50097 setCarryingCapacity Population 6915
event 50098 setCarryingCapacity Population 7004
event 50099 setCarryingCapacity Population 7094
event 50100 setCarryingCapacity Population 7185
event 50101 setCarryingCapacity Population 7278
event 50102 setCarryingCapacity Population 7372
event 50103 setCarryingCapacity Population 7467
event 50104 setCarryingCapacity Population 7563
event 50105 setCarryingCapacity Population 7660
event 50106 setCarryingCapacity Population 7759
event 50107 setCarryingCapacity Population 7859
event 50108 setCarryingCapacity Population 7960
event 50109 setCarryingCapacity Population 8063
event 50110 setCarryingCapacity Population 8167
event 50111 setCarryingCapacity Population 8272
event 50112 setCarryingCapacity Population 8379
event 50113 setCarryingCapacity Population 8487
event 50114 setCarryingCapacity Population 8596
event 50115 setCarryingCapacity Population 8707
event 50116 setCarryingCapacity Population 8819
event 50117 setCarryingCapacity Population 8933
event 50118 setCarryingCapacity Population 9048
event 50119 setCarryingCapacity Population 9165
event 50120 setCarryingCapacity Population 9283
event 50121 setCarryingCapacity Population 9403
event 50122 setCarryingCapacity Population 9524
event 50123 setCarryingCapacity Population 9647
event 50124 setCarryingCapacity Population 9771
event 50125 setCarryingCapacity Population 9897
event 50126 setCarryingCapacity Population 10024
event 50127 setCarryingCapacity Population 10153
event 50128 setCarryingCapacity Population 10284
event 50129 setCarryingCapacity Population 10416
event 50130 setCarryingCapacity Population 10550
event 50131 setCarryingCapacity Population 10686
event 50132 setCarryingCapacity Population 10824
event 50133 setCarryingCapacity Population 10963
```

```
event 50134 setCarryingCapacity Population 11104
event 50135 setCarryingCapacity Population 11247
event 50136 setCarryingCapacity Population 11392
event 50137 setCarryingCapacity Population 11539
event 50138 setCarryingCapacity Population 11688
event 50139 setCarryingCapacity Population 11839
event 50140 setCarryingCapacity Population 11991
event 50141 setCarryingCapacity Population 12145
event 50142 setCarryingCapacity Population 12301
event 50143 setCarryingCapacity Population 12459
event 50144 setCarryingCapacity Population 12619
event 50145 setCarryingCapacity Population 12782
event 50146 setCarryingCapacity Population 12947
event 50147 setCarryingCapacity Population 13114
event 50148 setCarryingCapacity Population 13283
event 50149 setCarryingCapacity Population 13454
event 50150 setCarryingCapacity Population 13627
event 50151 setCarryingCapacity Population 13803
event 50152 setCarryingCapacity Population 13981
event 50153 setCarryingCapacity Population 14161
event 50154 setCarryingCapacity Population 14343
event 50155 setCarryingCapacity Population 14528
event 50156 setCarryingCapacity Population 14715
event 50157 setCarryingCapacity Population 14905
event 50158 setCarryingCapacity Population 15097
event 50159 setCarryingCapacity Population 15291
event 50160 setCarryingCapacity Population 15488
event 50161 setCarryingCapacity Population 15688
event 50162 setCarryingCapacity Population 15890
event 50163 setCarryingCapacity Population 16095
event 50164 setCarryingCapacity Population 16302
event 50165 setCarryingCapacity Population 16512
event 50166 setCarryingCapacity Population 16725
event 50167 setCarryingCapacity Population 16940
event 50168 setCarryingCapacity Population 17158
event 50169 setCarryingCapacity Population 17379
event 50170 setCarryingCapacity Population 17603
event 50171 setCarryingCapacity Population 17830
event 50172 setCarryingCapacity Population 18060
event 50173 setCarryingCapacity Population 18293
event 50174 setCarryingCapacity Population 18529
event 50175 setCarryingCapacity Population 18768
event 50176 setCarryingCapacity Population 19010
event 50177 setCarryingCapacity Population 19255
event 50178 setCarryingCapacity Population 19503
event 50179 setCarryingCapacity Population 19754
event 50180 setCarryingCapacity Population 20008
event 50181 setCarryingCapacity Population 20266
event 50182 setCarryingCapacity Population 20527
event 50183 setCarryingCapacity Population 20791
event 50184 setCarryingCapacity Population 21059
event 50185 setCarryingCapacity Population 21330
event 50186 setCarryingCapacity Population 21605
event 50187 setCarryingCapacity Population 21883
event 50188 setCarryingCapacity Population 22165
event 50189 setCarryingCapacity Population 22451
event 50190 setCarryingCapacity Population 22740
```

```
event 50191 setCarryingCapacity Population 23033
event 50192 setCarryingCapacity Population 23330
event 50193 setCarryingCapacity Population 23631
event 50194 setCarryingCapacity Population 23935
event 50195 setCarryingCapacity Population 24243
event 50196 setCarryingCapacity Population 24555
event 50197 setCarryingCapacity Population 24871
event 50198 setCarryingCapacity Population 25191
event 50199 setCarryingCapacity Population 25515
event 50200 setCarryingCapacity Population 25844
event 50201 setCarryingCapacity Population 26177
event 50202 setCarryingCapacity Population 26514
event 50203 setCarryingCapacity Population 26856
event 50204 setCarryingCapacity Population 27202
event 50205 setCarryingCapacity Population 27552
event 50206 setCarryingCapacity Population 27907
event 50207 setCarryingCapacity Population 28266
event 50208 setCarryingCapacity Population 28630
event 50209 setCarryingCapacity Population 28999
event 50210 setCarryingCapacity Population 29373
event 50211 setCarryingCapacity Population 29751
event 50212 setCarryingCapacity Population 30134
event 50213 setCarryingCapacity Population 30522
event 50214 setCarryingCapacity Population 30915
event 50215 setCarryingCapacity Population 31313
event 50216 setCarryingCapacity Population 31716
event 50217 setCarryingCapacity Population 32125
event 50218 setCarryingCapacity Population 32539
event 50219 setCarryingCapacity Population 32958
event 50220 setCarryingCapacity Population 33383
event 50221 setCarryingCapacity Population 33813
event 50222 setCarryingCapacity Population 34249
event 50223 setCarryingCapacity Population 34690
event 50224 setCarryingCapacity Population 35137
event 50225 setCarryingCapacity Population 35590
event 50226 setCarryingCapacity Population 36048
event 50227 setCarryingCapacity Population 36512
event 50228 setCarryingCapacity Population 36982
event 50229 setCarryingCapacity Population 37458
event 50230 setCarryingCapacity Population 37940
event 50231 setCarryingCapacity Population 38429
event 50232 setCarryingCapacity Population 38924
event 50233 setCarryingCapacity Population 39425
event 50234 setCarryingCapacity Population 39933
event 50235 setCarryingCapacity Population 40447
event 50236 setCarryingCapacity Population 40968
event 50237 setCarryingCapacity Population 41496
event 50238 setCarryingCapacity Population 42031
event 50239 setCarryingCapacity Population 42572
event 50240 setCarryingCapacity Population 43120
event 50241 setCarryingCapacity Population 43675
event 50242 setCarryingCapacity Population 44238
event 50243 setCarryingCapacity Population 44808
event 50244 setCarryingCapacity Population 45385
event 50245 setCarryingCapacity Population 45970
event 50246 setCarryingCapacity Population 46562
event 50247 setCarryingCapacity Population 47162
```

```
event 50248 setCarryingCapacity Population 47769
event 50249 setCarryingCapacity Population 48384
event 50250 setCarryingCapacity Population 49007
event 50251 setCarryingCapacity Population 49638
event 50252 setCarryingCapacity Population 50277
event 50253 setCarryingCapacity Population 50925
event 50254 setCarryingCapacity Population 51581
event 50255 setCarryingCapacity Population 52245
event 50256 setCarryingCapacity Population 52918
event 50257 setCarryingCapacity Population 53600
event 50258 setCarryingCapacity Population 54290
event 50259 setCarryingCapacity Population 54989
event 50260 setCarryingCapacity Population 55697
event 50261 setCarryingCapacity Population 56414
event 50262 setCarryingCapacity Population 57141
event 50263 setCarryingCapacity Population 57877
event 50264 setCarryingCapacity Population 58623
event 50265 setCarryingCapacity Population 59378
event 50266 setCarryingCapacity Population 60143
event 50267 setCarryingCapacity Population 60918
event 50268 setCarryingCapacity Population 61703
event 50269 setCarryingCapacity Population 62498
event 50270 setCarryingCapacity Population 63303
event 50271 setCarryingCapacity Population 64118
event 50272 setCarryingCapacity Population 64944
event 50273 setCarryingCapacity Population 65781
event 50274 setCarryingCapacity Population 66628
event 50275 setCarryingCapacity Population 67486
event 50276 setCarryingCapacity Population 68355
event 50277 setCarryingCapacity Population 69235
event 50278 setCarryingCapacity Population 70127
event 50279 setCarryingCapacity Population 71030
event 50280 setCarryingCapacity Population 71945
event 50281 setCarryingCapacity Population 72872
event 50282 setCarryingCapacity Population 73811
event 50283 setCarryingCapacity Population 74762
event 50284 setCarryingCapacity Population 75725
event 50285 setCarryingCapacity Population 76700
event 50286 setCarryingCapacity Population 77688
event 50287 setCarryingCapacity Population 78689
event 50288 setCarryingCapacity Population 79703
event 50289 setCarryingCapacity Population 80730
event 50290 setCarryingCapacity Population 81770
event 50291 setCarryingCapacity Population 82823
event 50292 setCarryingCapacity Population 83890
event 50293 setCarryingCapacity Population 84971
event 50294 setCarryingCapacity Population 86066
event 50295 setCarryingCapacity Population 87175
event 50296 setCarryingCapacity Population 88298
event 50297 setCarryingCapacity Population 89435
event 50298 setCarryingCapacity Population 90587
event 50299 setCarryingCapacity Population 91754
event 50300 setCarryingCapacity Population 92936
event 50301 setCarryingCapacity Population 94133
event 50302 setCarryingCapacity Population 95346
event 50303 setCarryingCapacity Population 96574
event 50304 setCarryingCapacity Population 97818
```

```
event 50305 setCarryingCapacity Population 99078
event 50306 setCarryingCapacity Population 100354
event 50307 setCarryingCapacity Population 101647
event 50308 setCarryingCapacity Population 102956
event 50309 setCarryingCapacity Population 104282
event 50310 setCarryingCapacity Population 105625
event 50311 setCarryingCapacity Population 106986
event 50312 setCarryingCapacity Population 108364
event 50313 setCarryingCapacity Population 109760
event 50314 setCarryingCapacity Population 111174
event 50315 setCarryingCapacity Population 112606
event 50316 setCarryingCapacity Population 114056
event 50317 setCarryingCapacity Population 115525
event 50318 setCarryingCapacity Population 117013
event 50319 setCarryingCapacity Population 118520
event 50320 setCarryingCapacity Population 120047
event 50321 setCarryingCapacity Population 121593
event 50322 setCarryingCapacity Population 123159
event 50323 setCarryingCapacity Population 124745
event 50324 setCarryingCapacity Population 126352
event 50325 setCarryingCapacity Population 127980
event 50326 setCarryingCapacity Population 129629
event 50327 setCarryingCapacity Population 131299
event 50328 setCarryingCapacity Population 132990
event 50329 setCarryingCapacity Population 134703
event 50330 setCarryingCapacity Population 136438
event 50331 setCarryingCapacity Population 138195
event 50332 setCarryingCapacity Population 139975
event 50333 setCarryingCapacity Population 141778
event 50334 setCarryingCapacity Population 143604
event 50335 setCarryingCapacity Population 145454
event 50336 setCarryingCapacity Population 147328
event 50337 setCarryingCapacity Population 149226
event 50338 setCarryingCapacity Population 151148
event 50339 setCarryingCapacity Population 153095
event 50340 setCarryingCapacity Population 155067
event 50341 setCarryingCapacity Population 157064
event 50342 setCarryingCapacity Population 159087
event 50343 setCarryingCapacity Population 161136
event 50344 setCarryingCapacity Population 163212
event 50345 setCarryingCapacity Population 165314
event 50346 setCarryingCapacity Population 167443
event 50347 setCarryingCapacity Population 169600
event 50348 setCarryingCapacity Population 171785
event 50349 setCarryingCapacity Population 173998
event 50350 setCarryingCapacity Population 176239
event 50351 setCarryingCapacity Population 178509
event 50352 setCarryingCapacity Population 180808
event 50353 setCarryingCapacity Population 183137
event 50354 setCarryingCapacity Population 185496
event 50355 setCarryingCapacity Population 187885
event 50356 setCarryingCapacity Population 190305
event 50357 setCarryingCapacity Population 192756
event 50358 setCarryingCapacity Population 195239
event 50359 setCarryingCapacity Population 197754
event 50360 setCarryingCapacity Population 200301
event 50361 setCarryingCapacity Population 202881
```

```
event 50362 setCarryingCapacity Population 205494
event 50363 setCarryingCapacity Population 208141
event 50364 setCarryingCapacity Population 210822
event 50365 setCarryingCapacity Population 213538
event 50366 setCarryingCapacity Population 216289
event 50367 setCarryingCapacity Population 219075
event 50368 setCarryingCapacity Population 221897
event 50369 setCarryingCapacity Population 224755
event 50370 setCarryingCapacity Population 227650
event 50370 set maxOffspringNumber 2

output 50371
outputXML false
matingWithReplacement false

finalPedigreeDepth 2
extraPedigrees 0
extraSingleAscertainmentPedigrees 0
trackSNPs false
end

chromosome begin
length 250000

gene begin
name flank1
location 0
length 100000
gamma 0.316228 0.01
probabilityNoEffect 1
probabilityLethalEffect 0
probabilityPositiveEffect 0
end

gene begin
name exon1
location 100001
length 300
gamma 0.316228 0.01
probabilityNoEffect 0.44
probabilityLethalEffect 0
probabilityPositiveEffect 0
end

gene begin
name intron1
location 100303
length 3000
gamma 0.316228 0.01
probabilityNoEffect 1
probabilityLethalEffect 0
probabilityPositiveEffect 0
end

gene begin
name exon2
location 103305
```

```
length 300
gamma 0.316228 0.01
probabilityNoEffect 0.44
probabilityLethalEffect 0
probabilityPositiveEffect 0
end

gene begin
name intron2
location 103607
length 3000
gamma 0.316228 0.01
probabilityNoEffect 1
probabilityLethalEffect 0
probabilityPositiveEffect 0
end

gene begin
name exon3
location 106609
length 300
gamma 0.316228 0.01
probabilityNoEffect 0.44
probabilityLethalEffect 0
probabilityPositiveEffect 0
end

gene begin
name intron3
location 106911
length 3000
gamma 0.316228 0.01
probabilityNoEffect 1
probabilityLethalEffect 0
probabilityPositiveEffect 0
end

gene begin
name exon4
location 109913
length 300
gamma 0.316228 0.01
probabilityNoEffect 0.44
probabilityLethalEffect 0
probabilityPositiveEffect 0
end

gene begin
name intron4
location 110215
length 3000
gamma 0.316228 0.01
probabilityNoEffect 1
probabilityLethalEffect 0
probabilityPositiveEffect 0
end
```

```
gene begin
name exon5
location 113217
length 300
gamma 0.316228 0.01
probabilityNoEffect 0.44
probabilityLethalEffect 0
probabilityPositiveEffect 0
end

gene begin
name intron5
location 113519
length 3000
gamma 0.316228 0.01
probabilityNoEffect 1
probabilityLethalEffect 0
probabilityPositiveEffect 0
end

gene begin
name exon6
location 116521
length 300
gamma 0.316228 0.01
probabilityNoEffect 0.44
probabilityLethalEffect 0
probabilityPositiveEffect 0
end

gene begin
name intron6
location 116823
length 3000
gamma 0.316228 0.01
probabilityNoEffect 1
probabilityLethalEffect 0
probabilityPositiveEffect 0
end

gene begin
name exon7
location 119825
length 300
gamma 0.316228 0.01
probabilityNoEffect 0.44
probabilityLethalEffect 0
probabilityPositiveEffect 0
end

gene begin
name intron7
location 120127
length 3000
gamma 0.316228 0.01
probabilityNoEffect 1
probabilityLethalEffect 0
```

```
probabilityPositiveEffect 0
end

gene begin
name exon8
location 123129
length 300
gamma 0.316228 0.01
probabilityNoEffect 0.44
probabilityLethalEffect 0
probabilityPositiveEffect 0
end

gene begin
name flank2
location 123431
length 100000
gamma 0.316228 0.01
probabilityNoEffect 1
probabilityLethalEffect 0
probabilityPositiveEffect 0
end

end

phenotype begin
name total_selection
definition 1.0 + flank1 + exon1 + intron1 + exon2 + intron2 + exon3 + intron3
+ exon4 + intron4 + exon5 + intron5 + exon6 + intron6 + exon7 + intron7 +
exon8 + flank2
end

population begin
name Population
birth 0
initialSize 8100
carryingCapacity 8100
growthRate 10.0
death 60372
selection total_selection functional 1.0 * Phenotype
environmentNormal total_selection 0.0 0.0000001
familyEnvironmentNormal total_selection 0.0 0.0
end
```

Appendix A2

Comparing fine-mapping strategies for common SNPs at chr9p21

The manuscript and figures that follow were published in *Nature Genetics* in 2011. Supplementary material is not included below, but has been published online.

Comparing strategies to fine map the association of common SNPs on chromosome 9p21 to Type 2 Diabetes and Myocardial Infarction

Jessica Shea[1,2,3], Vineeta Agarwala[1,3,4,5], Anthony A. Philippakis[1,3,4,5,6,7], Jared Maguire[1], Eric Banks[1], Mark DePristo[1], Brian Thomson[1], Candace Guiducci[1], The Myocardial Infarction Genetics Consortium, Sekar Kathiresan[1,6,8,9,10], Stacey Gabriel[1], Noël P Burtt[1], Mark J. Daly[1,6,8,10], Leif Groop[11], and David Altshuler[1,3,6,9,10,12]

1. Program in Medical and Population Genetics, Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
2. Program in Biological and Biomedical Sciences, Harvard Medical School, Boston, Massachusetts, USA
3. Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA
4. Program in Biophysics, Harvard University, Cambridge, Massachusetts, USA
5. Harvard-Massachusetts Institute of Technology Division of Health Sciences and Technology, Harvard Medical School, Boston, Massachusetts, USA
6. Department of Medicine, Harvard Medical School, Boston, Massachusetts, USA
7. Department of Medicine, Brigham and Women's Hospital, Boston, Massachusetts, USA
8. Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts, USA
9. Cardiovascular Research Center, Massachusetts General Hospital, Massachusetts, USA
10. Department of Medicine, Massachusetts General Hospital, Boston, Massachusetts, USA
11. Department of Clinical Sciences, Diabetes and Endocrinology Research Unit, University Hospital Malmö, Lund University, Malmö, Sweden
12. Department of Molecular Biology, Massachusetts General Hospital, Boston, Massachusetts, USA

Non-coding variants at human chromosome 9p21 near *CDKN2A* and *CDKN2B* are associated with type 2 diabetes (T2D)[1-4], myocardial infarction (MI)[5-7], aneurysm[8], vertical cup disc ratio[9], and at least five cancers[10-16]. We compared approaches to more comprehensively assess genetic variation in the region. We performed targeted sequencing at high coverage in 47 individuals and compared the results to pilot data from the 1000 Genomes Project. We imputed variants into T2D and MI cohorts directly from targeted sequencing, from a genotyped reference panel derived from sequencing, and from 1000 Genomes low-coverage data. Common polymorphisms were captured similarly by all strategies. Imputation of intermediate frequency polymorphisms required a higher density of tag SNPs in disease

samples than available on first generation Genome Wide Association Study (GWAS) arrays. Association analyses identified more comprehensive sets of variants demonstrating equivalent statistical association to T2D or MI, but did not identify stronger associations the original GWAS signals.

Following the identification of a disease-associated region by GWAS, comprehensive study of sequence variation in the region is required to identify the full set of variants that might explain the association signal. Since GWAS arrays incompletely capture DNA variation in each region, it has been hypothesized that causal variants partially captured by linkage disequilibrium (LD) – due to location near recombination hotspots or lower minor allele frequency – might, if directly tested, display stronger association to phenotype than the tag SNPs used in GWAS. In particular, because HapMap and GWAS arrays contain primarily variants with minor allele frequency (MAF) >5%, first generation GWAS studies failed to test polymorphisms of somewhat lower frequency that might have larger effects on disease risk. Finally, even in regions where the true association signal is well captured by LD to array SNPs, enumeration of all associated variants is a necessary prerequisite to functional experiments that will identify causal mutation(s). Thus, an important next step following GWAS is to assemble a more complete catalog of variation present in an associated region, and to test it for association to the phenotype of interest.

With the advent of next generation sequencing and the emergence of data from the 1000 Genomes (1000G) Project, investigators must choose between (or combine) multiple strategies for creating and testing a reference panel of polymorphic sites. We re-sequenced ~240kb on chromosome 9p21 (chr9:21936711-22176221, hg18) spanning the T2D and MI associations in 47 unrelated individuals of European ancestry from the HapMap CEU population[17] as part of a sequencing project spanning six T2D-associated regions (**Supplementary Table A2.1**). Sequencing was performed at the Broad Institute on Illumina Genome Analyzers (**Supplementary Note A2,** all data available in the NCBI Short Read Archive). An analytical framework (**Supplementary Note A2, Supplementary Table A2.2, Supplementary Figs. A2.1-5**), since

extended and incorporated in the Genome Analysis Tool Kit[18,19], was developed and includes methods to empirically recalibrate Illumina base quality scores, a Bayesian framework to call SNPs, local re-alignment to identify insertions/deletions (and remove clusters of false positive SNPs), and filters to remove false positive SNP calls based on discrepancy between forward and reverse strands.

This targeted sequencing identified 635 high-confidence SNPs on chromosome 9p21 (4,463 across the six regions) (**Supplementary Table A2.3, Supplementary Fig. A2.6,** SNPs available in dbSNP). We evaluated sensitivity against HapMap II[17] and the high coverage Pilot 2 data from the 1000 Genomes Project[20] (**Supplementary Note A2**): at sites in overlapping samples with 10x or greater read coverage (70% of the region), sensitivity was 99% for HapMap variants and 97% for variants found in 1000G Pilot 2 (**Supplementary Fig. A2.7a-c**). To evaluate specificity, we genotyped 257 sites found on chromosome 9p21 but not previously genotyped in HapMap (**Supplementary Fig. A2.7d-e**). Overall, 96% of variants seen more than once in sequencing validated in the genotyping data (**Supplementary Table A2.4**).

We compared these variants to those discovered in the low-coverage Pilot 1 of the 1000 Genomes Project [20], limiting comparison to 32 CEU individuals studied in both projects. Across the six regions, both projects identified similar numbers of variants: 3,897 SNPs in the high coverage targeted sequencing as compared to 4,043 in 1000G Pilot 1. However, the variants found were in fact only partially overlapping. Of variants seen in the high coverage targeted sequencing, 22% were missed by 1000G Pilot 1 (**Fig. A2.1**), nearly all of which were rare: 72% of these sites were singletons and 12% were seen twice (**Fig. A2.1, Table A2.1**). Pilot 1 successfully identified 97% of SNPs seen more than 5 times in high coverage sequencing (**Table A2.1**). Of variants identified in Pilot 1 but not in targeted sequencing (n=998), nearly all were sites at which target capture failed to achieve high coverage: 65% of these sites had zero coverage. Thus, targeted capture and low-pass whole genome had distinct and non-overlapping failure modes.

We evaluated methods for testing these variants for association to disease via linkage

disequilibrium and haplotype-based imputation. First, we genotyped SNPs found in targeted re-sequencing on chromosome 9p21 in 168 individuals (56 parent offspring trios) from the HapMap extended CEU population[21] (**Supplementary Note A2**). We used MACH[22,23] to impute variants from this reference panel into 1,000 T2D patients and 1,048 controls from the Diabetes Genetics Initiative (DGI) cohort[1] and 1,274 MI cases and 1,407 controls from the Myocardial Infarction Genetics (MIGen) Consortium cohort[6], each previously genotyped on Affymetrix GWAS arrays (**Supplementary Note A2**).

Table A2.1: **Sensitivity of 1000 Genomes Pilot 1 for variants detected in targeted, high-coverage sequencing of samples common to both projects.**

Number of times non-reference allele observed in this study	Number of SNPs called, this study	% Contained in 1000G Pilot 1	% in dbSNP, build 129	% Validated on chr9p21
1X	941	35%	13%	91%
2X	300	68%	42%	88%
3X	239	82%	55%	100%
4X	154	87%	66%	86%
5X	186	91%	67%	70%
>5X	2077	97%	92%	98%

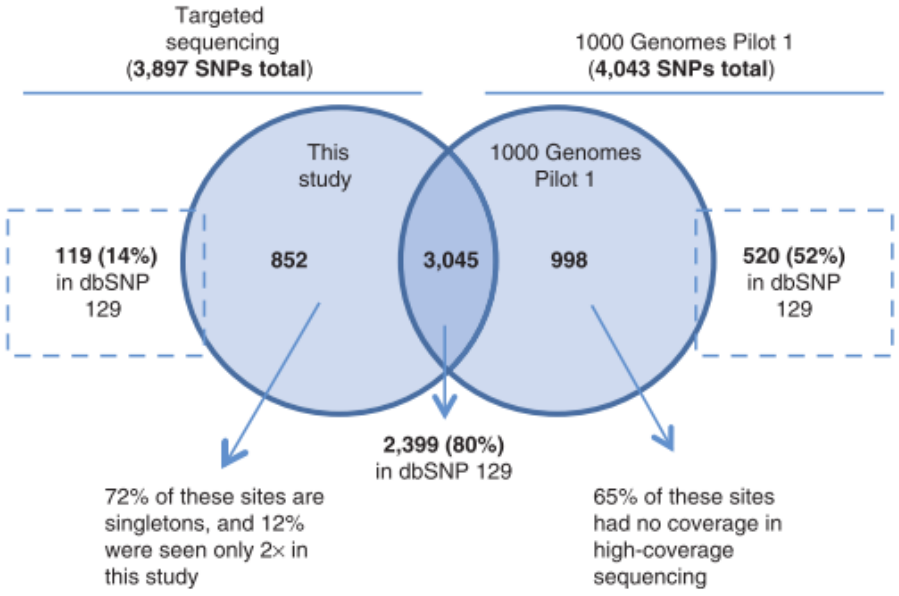


Figure A2.1: **Comparison of targeted sequencing to 1000 Genomes Pilot 1 data.** Variant calls were made in all six regions of type 2 diabetes association in the 32 individuals sequenced as part of both this targeted, high-coverage sequencing effort (47 CEU HapMap individuals) and 1000 Genomes Pilot 1 (60 CEU HapMap individuals).

We compared the results of imputation with this augmented reference panel (n=464 variants, **Supplementary Table A2.5**) to those obtained when imputing from HapMap II alone (n=238

variants). The addition of genotype data for a more complete collection of common variants provided imputation data for a much larger number of SNPs than was possible with HapMap II, which contains only 50-60% of common variants (**Fig. A2.2a-b, Supplementary Fig. A2.8a-b**). However, even with the augmented reference panel, the tag SNP density characteristic of the first generation GWAS arrays on which our disease samples were typed allowed only 80% of common (MAF > 5%) variants to be captured (either directly typed or imputed with a MACH-estimated $r^2 \geq 0.8$). Moreover, only a small fraction of intermediate frequency variation (MAF 2-5%) was imputed with an estimated r^2 above this stringent threshold (**Fig. A2.2c-d, Supplementary Fig. A2.8c-d**).

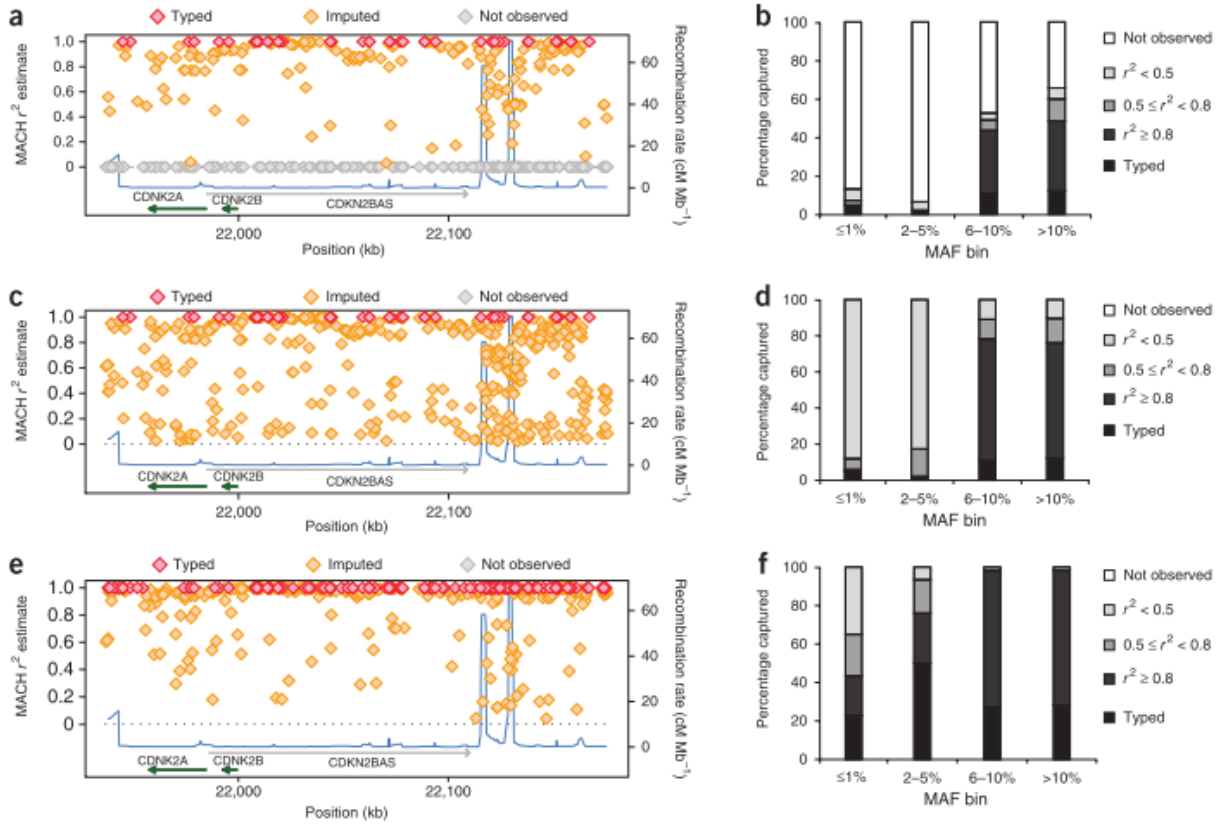


Figure A2.2: **Fraction of variation on chr 9p21 captured in T2D cohort by different imputation scenarios**
MACH imputation quality estimates **(a, c, e)** and overall fraction of variation captured in T2D samples **(b, d, f)** for different imputation scenarios. **(a, c, e)** The MACH-estimated r^2 for each SNP is plotted as a function of genomic position. SNPs not observed in the reference panel are assigned an r^2 of zero. Recombination rate (estimated from HapMap) is plotted to reflect local LD structure. Gene annotations were taken from the University of California-Santa Cruz Genome Browser. **(b, d, f)** The fraction of variants captured in T2D samples is shown as a function of MAF and MACH-estimated r^2. Imputation scenarios are: **(a, b)** Imputing from HapMap II (n=238 SNPs in 60 individuals) into the SNPs genotyped on the Affymetrix 500K array; **(c, d)** Imputing from 112 individuals genotyped at HapMap II sites and validated sequencing sites (total n=464 SNPs) into the SNPs genotyped on the Affymetrix 500K array; **(e, f)** Imputing from the same reference panel as c, d into SNPs genotyped on the Affymetrix 500K array plus additional tag SNPs genotyped in the T2D cohort (genotyped marker density in T2D samples ~1 SNP/1.5kb).

To evaluate the impact of tag SNP density on imputation performance, we increased the number of tags across the region to approximately 1 SNP per 1.5kb (the previous density was ~1SNP/5kb in T2D samples and ~1SNP/3kb in MI samples) in the T2D and MI cohorts (**Supplementary Note A2**). With this increased density of tag SNPs, nearly all common variants (~98%) were captured with $r^2 \geq 0.8$ in disease samples. Moreover, performance for intermediate frequency variants was dramatically improved, rising from 2% to 75% with $r^2 \geq 0.8$ (**Fig. A2.2e-f, Supplementary Fig. A2.8e-f**). This result was not specific to the Affymetrix GWAS arrays, as we observed a similar improvement in imputation ability upon addition of tag SNPs using multiple other GWAS arrays (**Supplementary Fig. A2.9**).

We next compared different reference panels, imputing in each case into disease samples with the higher tag SNP density. The reference panels were: (a) the genotyped reference panel of 168 individuals above (112 unrelated individuals), (b) the targeted sequencing data (47 individuals, without genotyping and expansion into a larger sample set), and (c) 1000 Genomes Pilot 1 (55 individuals). We considered both the fraction of variants in each reference panel successfully imputed (which is related to the quality and completeness of SNP genotypes and to the size of the reference panel) and the fraction of all variation captured (which, in addition, depends on the proportion of all known SNPs represented in the reference panel).

The union of the three reference panels contained 582 variants (**Fig. A2.3a**). Each panel was partially incomplete, due to genotyping assay failure in the genotyped panel (14% of SNPs missing), sample size and low coverage in 1000 genomes (16% of SNPs missing), and sample size and gaps in coverage in the targeted sequencing (19% of SNPs missing). For common variants, there is little difference in bulk performance between the reference panels. Considering only SNPs contained in each reference panel (**Fig. A2.3b**) the genotyped panel has the highest proportion of variants imputed well. However, when all variation is considered (**Fig. A2.3c**), a *lower* proportion of common variation is captured by imputing from the genotyped reference panel, owing to the fact that some SNPs were missing in this panel because they failed assay design or quality control. Notably,

the 1000G (freely available) and sequencing (costly) strategies performed equivalently for these common variants.

For intermediate frequency variants, there are more pronounced differences between the panels (**Fig. A2.3b-c**). These variants were best imputed from the genotyped reference panel (**Fig. A2.3b**), which was the largest and also contained trio information. This was true even when all variation was considered (**Fig. A2.3c**), suggesting that the improved imputation quality from genotype data and increased sample size offset the loss of variants in this panel due to genotyping failure. Comparing the high coverage re-sequencing and 1000G reference panels, lower frequency variants were better imputed from the high coverage re-sequencing data both when considering only the SNPs within each reference panel (**Fig. A2.3b**) and when considering the overall proportion of low frequency variants captured by imputation from each reference panel (**Fig. A2.3c**). This is consistent with the low coverage 1000G pilot 1 data being less complete and accurate for lower frequency variants[20].

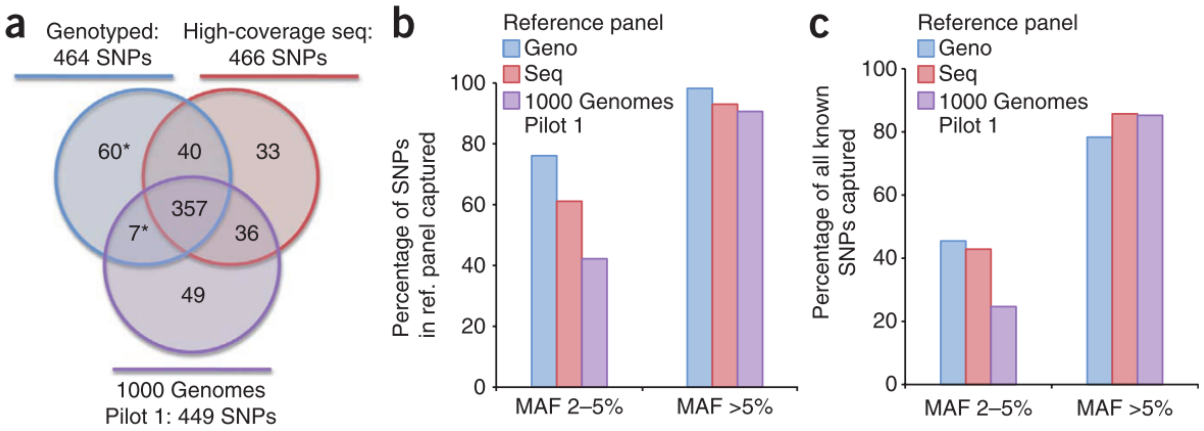


Figure A2.3: **Comparison of imputation from a genotyped reference panel, directly from high coverage re-sequencing data, and directly from 1000G Pilot 1 data**
(a) Variants present in the three reference panels and their overlap. The 67 variants present in the genotyped reference panel but not in the high coverage sequencing reference panel (denoted by asterisk) were called in high coverage sequencing as singletons and so were excluded from the sequencing reference panel. 40% of these variants are not singletons in the larger genotyped reference panel. **(b)** The fraction of sites within each reference panel captured with a MACH-estimated r^2 of at least 0.8. **(c)** The overall fraction of known variants captured with a MACH-estimated r^2 of at least 0.8 by imputation from each reference panel.

We tested variants for association to T2D and MI using imputation from all three reference panels to maximize the number of variants captured (**Supplementary Note A2**). Overall, we have captured 461 of the 582 polymorphic variants observed across all three reference panels in our T2D and MI samples with a MACH-estimated r^2 of at least 0.8: this represents ~ 92% of all known common variants and ~52% of intermediate frequency variants (at a MACH-estimated r^2 of 0.5, these figures are 98% of common variants and 83% of intermediate frequency variants). In comparison, only 176 of the 582 variants were previously captured by imputation from HapMap. Despite testing many additional SNPs in partial LD with the index GWAS hits and at allele frequencies not well captured by first generation GWAS arrays and HapMap, we found no example of a SNP with stronger association to T2D or MI than the initial GWAS signals.

However, we did identify multiple additional variants in strong LD with the GWAS hits that might underlie each association. We observed the three-tiered haplotypic association to T2D first reported by the Wellcome Trust Case Control Consortium with protective, risk, and neutral haplotypes (**Table A2.2**). The protective alleles of the GWAS SNP (rs10811661) and nine other SNPs in strong LD with this variant tag the protective haplotype (**Fig. A2.4a, Supplementary Table A2.6**). Interestingly, no single SNP yet identified marks the risk haplotype. Association analyses for MI identified 7 SNPs in LD with each other and with equivalent evidence for association ($P < 10^{-4}$) as well as 54 additional SNPs with only slightly less evidence for association ($P < 10^{-3}$) (**Fig. A2.4b, Supplementary Table A2.6**). Knockout of the MI-associated region in mouse alters regulation of *CDKN2A* and *CDKN2B*[24], and two of the associated SNPs have recently been shown to disrupt a STAT1 binding site[25]. Interestingly, in addition to the SNPs disrupting the STAT1 site, there are other variants with equivalent MI association and with putative functional annotations, including variants overlapping exons of the non-coding transcript *CDKN2BAS*, highly conserved regions, and predicted, conserved transcription factor binding sites (**Supplementary Table A2.6**).

Table A2.2: Haplotypic association to T2D on chromosome 9p21

rs10757282 and the reported SNP from GWAS, rs10811661, define haplotypes with three levels of risk (risk, protective, and neutral) for T2D.

Haplotypes defined by rs10757282, rs10811661			
Haplotype	**Frequency**	**OR**	***P*-value**
Overall Evidence	--	--	4.40×10^{-5}
CT	0.30	1.29	3.99×10^{-4}
TT	0.54	0.96	5.24×10^{-1}
CC	0.16	0.72	2.71×10^{-4}

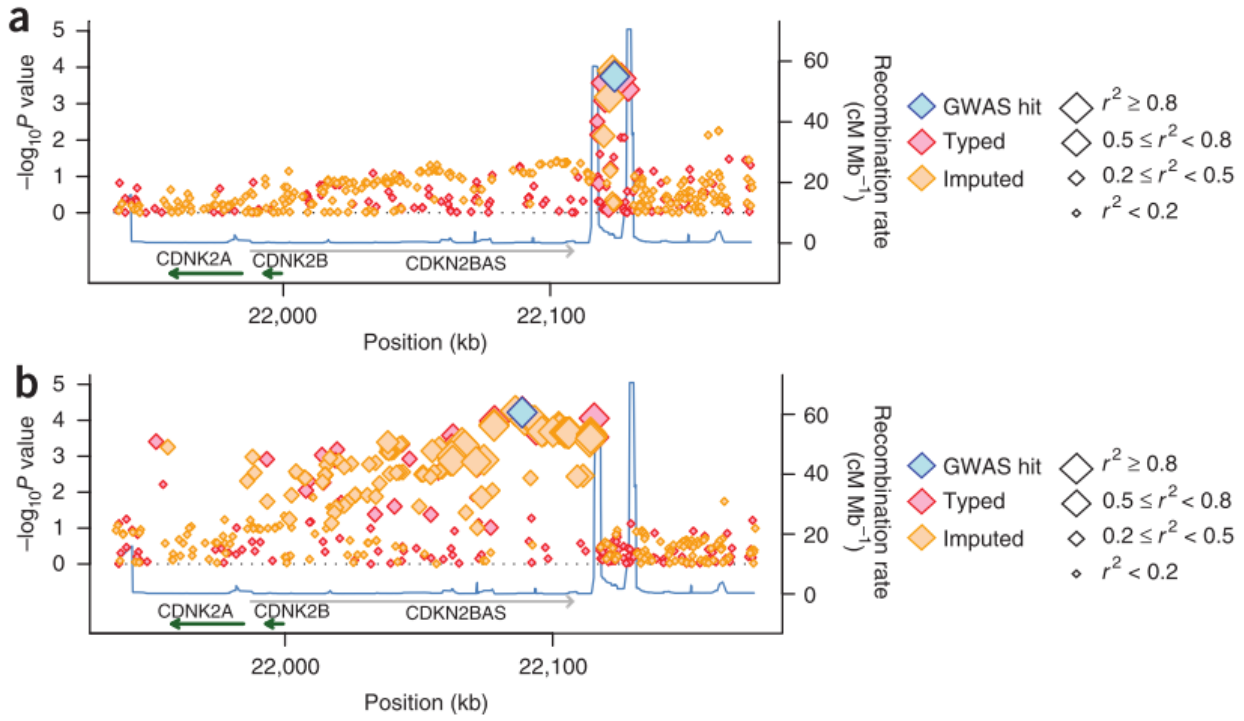


Figure A2.4: Association results for T2D and MI on chromosome 9p21

Regional plots showing association signal for **(a)** T2D and **(b)** MI. The signal for each SNP (represented as -$\log_{10}$ p-value) is plotted as a function of genomic position. The size of the diamond for each SNP represents the LD (measured as r^2) between that SNP and the original GWAS SNP (rs10811661 for T2D and rs4977574 for MI). Recombination rate (estimated from HapMap) is plotted to reflect the local LD structure in the region. Gene annotations were taken from the University of California-Santa Cruz Genome Browser.

This study is limited by the investigation of a single region (albeit one with at least eight different disease associations), by the early nature of the sequencing data analyzed, by the small number of samples sequenced in SNP discovery, and by the sample size of our disease cohorts. Nonetheless, the observations on the strengths and weaknesses of different methods for fine mapping GWAS signals are likely general: targeted high coverage sequencing provides high

sensitivity for lower frequency variants, but has gaps in coverage; the 1000G Pilot 1 resource offers more even coverage at lower depth, currently sufficient for capture of most common variation; creating a genotyped reference panel improves accuracy and sample size, but is limited by assay conversion failures; tag SNP density characteristic of first generation GWAS is inadequate to maximally extract information with current imputation algorithms. To some extent, these limitations are transient: the growing 1000 Genomes Project resource is sequencing over 2,000 diverse samples with both low-pass whole genome and high coverage targeted exon approaches, increasing the accuracy and completeness of the public reference panel. However, our results suggest that fully exploiting this resource for imputation may require increasing tag density in GWAS disease samples and / or improved algorithms for imputation.

Finally, our study did not find evidence for stronger association at 9p21 to SNPs in moderate LD with the initial tags. While the maximum achievable association signal for lower frequency variants was limited by our sample size, we did not observe lower frequency variants with effect sizes that could individually explain the common variant associations. We do, however, identify additional common variants in LD with the GWAS hits that might underlie each association. Enumeration of all variants on 9p21 that might explain each association signal will be needed as a foundation for systematic functional studies that aim to understand how different non-coding variants in this single genomic interval can lead to such varied and clinically significant phenotypic associations.

Methods

Targeted Re-Sequencing

Six regions associated with T2D were selected for targeted re-sequencing (**Supplementary Table A2.1**). Because the goal of this study to was to identify additional SNPs that might explain the initial GWAS signal, region boundaries were selected to encompass all SNPs showing detectable linkage disequilibrium ($r^2 \geq 0.2$) to the T2D associated SNP with the most significant p-value. DNA was captured for sequencing by long-range PCR with 2-5kb amplicons or by hybrid selection (HS) using 170bp baits tiled across the region on an Agilent microarray[26]. All sequencing was performed at the Broad Institute in 2008 using Illumina Genome Analyzers. Runs from PCR-based capture generated 36bp reads and runs from HS-based capture generated 46-50bp reads. Methods for alignment,

quality score adaptation and recalibration, and variant calling are described in detail in the **Supplementary Note A2.**

SNP Genotyping and Quality Control

Genotyping was performed on the Sequenom MassARRAY iPLEX platform. Quality control filters included 1) > 95% genotyping rate, 2) Hardy Weinberg equilibrium (with $P > 0.001$) and 3) Mendel error rate < 5%.

Phasing and Imputation

We compared several strategies and publicly available tools for phasing and imputation directly from Illumina sequencing data (**Supplementary Fig. A2.10-11**). Phased haplotypes for all reference panels were created using the PHASE software package (Version 2.1)[27,28]. For the genotyped reference panel, trio information was used in phasing (-P1 option). For sequencing reference panels, known phase was specified at HapMap sites (-k option). All other PHASE parameters were default values. Imputation from reference haplotypes was performed using MACH[22,23] (Version 1.0.16). 100 rounds were used; all other MACH parameters were default values.

Association Analyses

Variants were tested for association using logistic regression on imputed genotype dosages and individual disease status. EIGENSTRAT[29] (DGI) or PLINK[30] (MIGen) was used to estimate principal components which track with the ancestry of the study samples[1,6]; the first ten components were used as covariates in logistic regression to account for population structure. For T2D analyses, additional covariates used were: age, gender, and body mass index. For MI analyses additional covariates used were age, gender, BMI, and smoking. Tests for haplotypic association to T2D were performed using the PLINK[30] (Version 1.05) software package.

Acknowledgements

Patient collections in the DGI study were funded by grants from the Sigrid Juselius and Folkhälsan foundations as well as the Swedish Research Council (LG). The DGI GWAS study was supported by a grant from Novartis.

The MIGen study was funded by the US National Institutes of Health (NIH) and National Heart, Lung, and Blood Institute's STAMPEED genomics research program through a grant to D.A (R01 HL087676). S.K. is supported by a Doris Duke Charitable Foundation Clinical Scientist Development Award, a charitable gift from the Fannie E. Rippel Foundation, the Donovan Family Foundation, a career development award from the NIH and the Department of Medicine and Cardiovascular Research Center at Massachusetts General Hospital. DA and JS are supported in part by a Distinguished Clinical Scholar Award from the Doris Duke Charitable Foundation (to D.A.)

Next-generation sequencing for this work was performed by the Broad Institute Sequencing Platform and genotyping was performed by the Broad Institute Genetic Analysis Platform. We acknowledge their excellence and collaboration on this study. Sequencing was supported in part by a grant from NHGRI and by the Broad Institute.

The authors thank Manny Rivas, Andrey Sivachenco, and Kiran Garimella for helpful discussions on sequencing and Benjamin Voight, Stephan Ripke, and Ron Do for helpful discussions on imputation.

Author Contributions

Manuscript writing: J.S., V.A., A. A. P., D.A.

Clinical samples: S.K., L.G., D.A., The Myocardial Infarction Genetics Consortium

Next-generation sequencing data generation: C.G., N.P.B., S.G., Broad Institute Sequencing Platform

Sequencing analysis and variant calling: A.A.P, J.M., E.B., M.D., S.G., M.J.D., D.A.

Imputation and association analysis: J.S., V.A., M.J.D., D.A.

Genotyping and analysis: J.S., V.A, B.T., C.G., N.P.B., Broad Institute Genetic Analysis Platform

References

1. Saxena, R.*, et al.* Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* **316**, 1331-1336 (2007).
2. Scott, L.J.*, et al.* A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* **316**, 1341-1345 (2007).
3. Zeggini, E.*, et al.* Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat Genet* **40**, 638-645 (2008).
4. Zeggini, E.*, et al.* Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* **316**, 1336-1341 (2007).
5. Helgadottir, A.*, et al.* A common variant on chromosome 9p21 affects the risk of myocardial infarction. *Science* **316**, 1491-1493 (2007).
6. Kathiresan, S.*, et al.* Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. *Nat Genet* **41**, 334-341 (2009).
7. McPherson, R.*, et al.* A common allele on chromosome 9 associated with coronary heart disease. *Science* **316**, 1488-1491 (2007).
8. Helgadottir, A.*, et al.* The same sequence variant on 9p21 associates with myocardial infarction, abdominal aortic aneurysm and intracranial aneurysm. *Nat Genet* **40**, 217-224 (2008).
9. Ramdas, W.D.*, et al.* A genome-wide association study of optic disc parameters. *PLoS Genet* **6**, e1000978 (2010).
10. Bishop, D.T.*, et al.* Genome-wide association study identifies three loci associated with melanoma risk. *Nat Genet* **41**, 920-925 (2009).
11. Falchi, M.*, et al.* Genome-wide association study identifies variants at 9p21 and 22q13 associated with development of cutaneous nevi. *Nat Genet* **41**, 915-919 (2009).
12. Sherborne, A.L.*, et al.* Variation in CDKN2A at 9p21.3 influences childhood acute lymphoblastic leukemia risk. *Nat Genet* **42**, 492-494.
13. Shete, S.*, et al.* Genome-wide association study identifies five susceptibility loci for glioma. *Nat Genet* **41**, 899-904 (2009).
14. Stacey, S.N.*, et al.* New common variants affecting susceptibility to basal cell carcinoma. *Nat Genet* **41**, 909-914 (2009).
15. Turnbull, C.*, et al.* Genome-wide association study identifies five new breast cancer susceptibility loci. *Nat Genet* **42**, 504-507.
16. Wrensch, M.*, et al.* Variants in the CDKN2B and RTEL1 regions are associated with high-grade glioma susceptibility. *Nat Genet* **41**, 905-908 (2009).
17. Frazer, K.A.*, et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851-861 (2007).
18. Depristo, M.A.*, et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**, 491-498 (2011).

19. McKenna, A.*, et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303 (2010).
20. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073 (2010).
21. International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52-58 (2010).
22. Li, Y.a.A.G. MACH 1.0: rapid haplotype reconstructoin and missing genotype inference. *American Journal of Human Genetics* **S70**, 2290 (2006).
23. Li, Y., Willer, C., Sanna, S. & Abecasis, G. Genotype imputation. *Annu Rev Genomics Hum Genet* **10**, 387-406 (2009).
24. Visel, A.*, et al.* Targeted deletion of the 9p21 non-coding coronary artery disease risk interval in mice. *Nature* **464**, 409-412.
25. Harismendy, O.*, et al.* 9p21 DNA variants associated with coronary artery disease impair interferon-gamma signalling response. *Nature* **470**, 264-268 (2011).
26. Gnirke, A.*, et al.* Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol* **27**, 182-189 (2009).
27. Stephens, M. & Scheet, P. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am J Hum Genet* **76**, 449-462 (2005).
28. Stephens, M., Smith, N.J. & Donnelly, P. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* **68**, 978-989 (2001).
29. Price, A.L.*, et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904-909 (2006).
30. Purcell, S.*, et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-575 (2007).

Appendix A3

Exploring the role of 3D genome structure in human translocations

The manuscript and figures that follow were published in *PLoS One* in 2012. Supplementary material is not included in entirety below, but has been published online.

Three-dimensional genome structure influences partner selection for chromosomal translocations

Jesse M. Engreitz[1,2*], Vineeta Agarwala[1,2,3*], and Leonid A. Mirny[1,4^]

[1]Harvard-MIT Division of Health Sciences and Technology, MIT, Cambridge, MA 02139, USA
[2]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
[3]Biophysics Program, Graduate School of Arts and Sciences, Harvard University, Cambridge, MA 02138, USA
[4]Department of Physics, MIT, Cambridge, MA 02139, USA
[*]Equal contribution
^E-mail: leonid@mit.edu

Abstract

The human genome adopts nonrandom three-dimensional conformations within the nucleus. It is unclear, however, the extent to which genome structure contributes to chromosomal rearrangements in human disease. In particular, spatial co-localization of chromosomal breakpoints prior to rearrangement has been implicated in the genesis of chromosomal translocations in cancer. We performed a genome-wide analysis by intersecting Hi-C spatial proximity maps with collections of 1,533 chromosomal translocations from cancer and germline genomes. Hi-C detects spatial co-localization of many translocation partners in normal human cells prior to rearrangement, including the cancer translocation partners *BCR-ABL* and *MYC-IGH*. While translocation breakpoints occur most commonly in broad regions of open chromatin, we demonstrate that long-range physical interactions between pairs of genomic loci additionally predispose them to rearrange. Moreover, translocation breakpoints reported in human hematologic malignancies are particularly enriched for

co-localization signal in chromosome conformation experiments conducted in a lymphoid cell line, suggesting that genome structure may provide one mechanism for tissue-specific chromosomal alterations. Hi-C also reveals existing chromosomal rearrangements in abnormal genomes, allowing for detection and fine-mapping of novel chromosomal translocations in the K562 erythroleukemic cell line. Our results support a broad role for three-dimensional genome structure in translocation-partner selection and establish Hi-C as a key method for dissecting the structural features that contribute to human disease.

Author Summary

Chromosomal translocations are key features of cancer genomes that contribute to disease progression. These alterations involve breaks in the primary DNA sequence followed by illegitimate repair between inappropriate pairs of genomic loci. While the functional consequences of these mutations can be severe, the factors that predispose certain genomic regions to rearrange with each other are poorly understood. One attractive hypothesis to explain translocation partner preferences is that pairs of rearrangement-prone loci physically co-localize in the nuclei of normal cells; when one partner is damaged, it is mistakenly joined to a nearby available sequence. While individual examples support this prediction, a genome-wide demonstration of this principle has proven elusive due to limitations in the scale and resolution of microscopy. In this study, we leverage recent advances in sequence-based interrogation of three-dimensional genome conformation to show, for the first time, that many chromosomal translocation partners co-localize prior to rearrangement. We further find that the same technology can precisely locate breakpoints of rearranged loci after the mutagenic event. This insight supports an important role for three-dimensional genome architecture in the genesis of cancer translocations.

Introduction

Chromosomal translocations play an important role in both inherited and acquired human disease. Translocations affect cellular function by changing gene copy number, creating fusion genes with aberrant function, or repositioning regulatory elements. In cancer, translocations

contribute to malignant transformation of hematologic cells [1, 2] and have more recently been implicated in multiple epithelial neoplasms [3, 4, 5, 6]. Classic examples of driver rearrangements in cancer are the *BCR-ABL* translocation (observed in >90% of cases of chronic myeloid leukemia) and the *TMPRSS2-ERG* fusion (observed in ~50% of prostate cancers) [7, 8, 9, 10].

Emerging evidence suggests that the formation of chromosomal rearrangements is a nonrandom process. Repeated observation of specific translocations, as well as the existence of rearrangement hotspots in cancer [11], suggests that intrinsic cellular and genomic features predispose certain regions to translocate. One attractive hypothesis is that higher-order genome organization - that is, the physical proximity of chromosomes in the nucleus *prior* to translocation - contributes to the occurrence of specific translocations [12, 13].

Indeed, recent work demonstrates that the human genome adopts nonrandom conformations in the nucleus [14, 15, 16, 17, 18], suggesting that three-dimensional genome architecture could play a role in translocation partner selection. Case studies using fluorescence *in situ* hybridization (FISH) show that genes involved in recurrent translocations in several cancer types are positioned non-randomly and relatively close to one another in the nuclei of normal cells prior to malignant transformation [5, 19, 20, 21]. Furthermore, the intermingling of chromosomal territories appears to correlate with translocation frequency between pairs of chromosomes [22]. These data suggest that physical interaction between non-homologous chromosomes may be an important mechanism underlying recurrent translocations. However, current imaging methods have insufficient resolution and throughput to examine these relationships on a large scale. Without a genome-wide, high-resolution technology to characterize spatial organization, we have limited ability to determine whether proximity contributes to the thousands of observed chromosomal rearrangements in cancer. The physical size of the genomic regions involved in these interactions, as well as their tissue-specificity, are also uncharacterized.

Here we leverage Hi-C, a next-generation sequencing method for probing the three-dimensional architecture of the genome [18], to investigate the structural features that may

contribute to translocation partner preferences. We systematically test the hypothesis that translocations occur between spatially co-localized regions of the genome, integrating a total of 1,533 chromosomal rearrangements from both cytogenetic and sequencing-derived datasets. We find that many translocation partners are located in broad chromatin domains that interact in normal cells, thus predisposing them to chromosomal rearrangements. Hi-C also identifies existing rearrangements in malignant cells and enables fine-mapping of chromosomal breakpoints. Our results support a broad role for 3D genome structure in translocation-partner selection and establish Hi-C as a method for dissecting structural features contributing to human disease.

Results

Hi-C detects proximity of canonical translocation partners. We obtained Hi-C interaction maps generated in lymphoblastoid and erythroleukemic cell lines from Liebermann et al. [18]. Previous iterations of chromosome conformation capture technology have succeeded in identifying specific interactions between genes and regulatory elements [23, 24] as well as between actively transcribed or repressed genes [25, 26]. However, it was unclear whether Hi-C would have the power to detect the variable interactions that occur between translocation partners in an aggregate cell population.

To address this question, we examined the canonical translocation partners *BCR* and *ABL*, which form an unbalanced, often-amplified rearrangement in K562 cells (**Figure A3.1A**) and have been shown by FISH to co-localize in the nuclei of multiple normal hematopoietic cell types prior to the translocation event [12, 19, 27]. In K562 cells, Hi-C detected a strong interaction (1,996 reads) in the 1-Mb bin containing *BCR-ABL* (**Figure A3.1C**). In immortalized lymphoblastoid cells that do not harbor the translocation (**Figure A3.1B**), on the other hand, the signal in the region surrounding *BCR-ABL* was markedly reduced (10 reads, **Figure A3.1D**). The large difference in raw read count between the two cell lines reflects the fact that these regions of chromosomes and 9 and 22 interact in *cis* in K562 and *trans* in GM06690.

While this analysis highlighted the existing *BCR-ABL* translocation in K562 cells, the raw read count data are susceptible to biases introduced by differences in mappability and frequency of restriction enzyme sites in each megabase bin. To more robustly examine the weaker interactions between loci prior to rearrangement, we calculated a normalized Hi-C score (see Methods) for each megabase bin and again examined the *BCR-ABL* in K562 (**Figure A3.1E**) and GM06690 (**Figure A3.1F**) cells. Remarkably, in the karyotypically normal GM06690 cell line, the interaction extended beyond the megabase including *BCR* and *ABL* and included much of 9q34 and 22q11 (**Figure A3.1F**). To quantify this relationship, we calculated the average interaction score across the two-dimensional interaction map for 9q34 and 22q11. Compared to random regions of the same size on chromosomes 9 and 22, the bands containing *BCR* and *ABL* fell above the 90th and 95th percentiles of interaction scores in K562 and GM06690 cells, respectively (**Figure A3.1G**). The relatively lower score for the K562 cell line in the normalized data is due to the correction for the total number of *trans*-chromosomal reads observed for each megabase.

We next examined the loci involved in the t(8;14)(q24;q32) translocation, a rearrangement associated with Burkitt's lymphoma that brings the oncogene *MYC* under the control of activating enhancer elements at the *IGH* locus [28]. In lymphoblastoid cells, the chromosomal bands including *MYC* and *IGH* interacted more strongly than regions of the same size on the same chromosome pair, thus representing a local hotspot of *trans*-interaction between chromosomes 8 and 14 (**Figure A3.1H**). In comparison, the *trans*-interactions between the control loci *MYC-TGFBR2* and *IGH-TGFBR2*, which are not observed to undergo translocation in cancer cells, did not differ significantly from background, consistent with previous results obtained with FISH [20]. These results demonstrated that Hi-C can detect the spatial co-localization of translocation partners in karyotypically normal cells.

Many translocation partners co-localize in the nucleus. To assess the broader role of three-dimensional genome architecture in translocation partner selection, we gathered four datasets totaling 1,533 chromosomal rearrangements (**Table A3.1**) [29, 30, 31]. Identified by cytogenetic

and high-throughput sequencing modalities in cancer and germline genomes from multiple tissues, these four genome-wide datasets broadly sampled the space of possible chromosomal translocations (see Supplemental Materials).

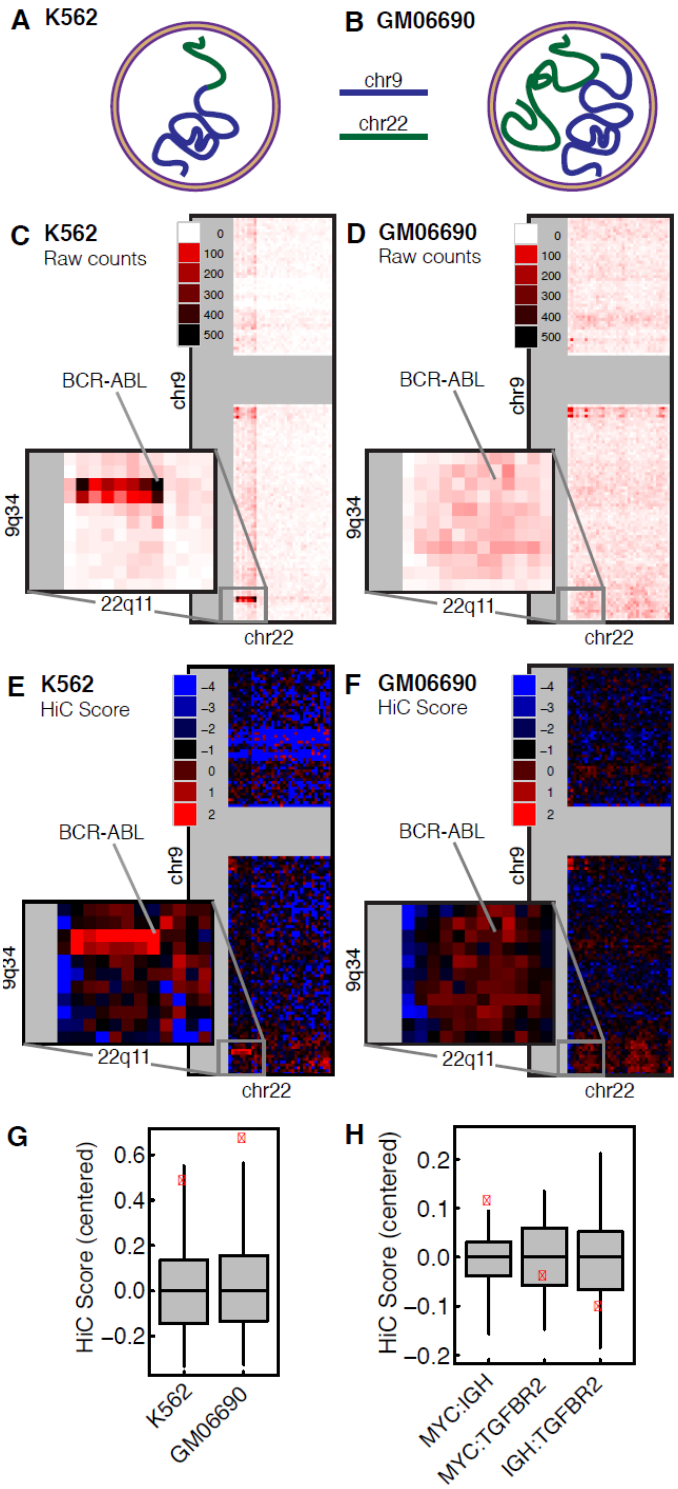


Figure A3.1: Hi-C detects interaction between known translocation partners BCR-ABL and MYC-IGH.

Chromosomes 9 and 22 are physically joined in **(A)** K562, but not in **(B)** GM06690. Hi-C maps show the extremely high read counts mapping to the megabase bin containing *BCR* (22q11) and *ABL* (9q34) in **(C)** K562 compared to **(D)** GM06690. **(E and F)** Normalized Hi-C scores provide better resolution for detecting *trans*-chromosomal interactions. **(G)** Centered interaction scores of 9q34:22q11 (red dots) are compared to the distribution of scores for random regions of the same size from chromosomes 9 and 22, excluding centromeres. **(H)** Centered interaction scores for the chromosomal bands containing the translocation partners *MYC-IGH* as well as the control partners *MYC-TGFBR2* and *IGH-TGFBR2*, compared to the background distribution of scores on the same chromosome pairs. Error bars in **(G)** and **(H)** represent the 5th and 95th percentiles.

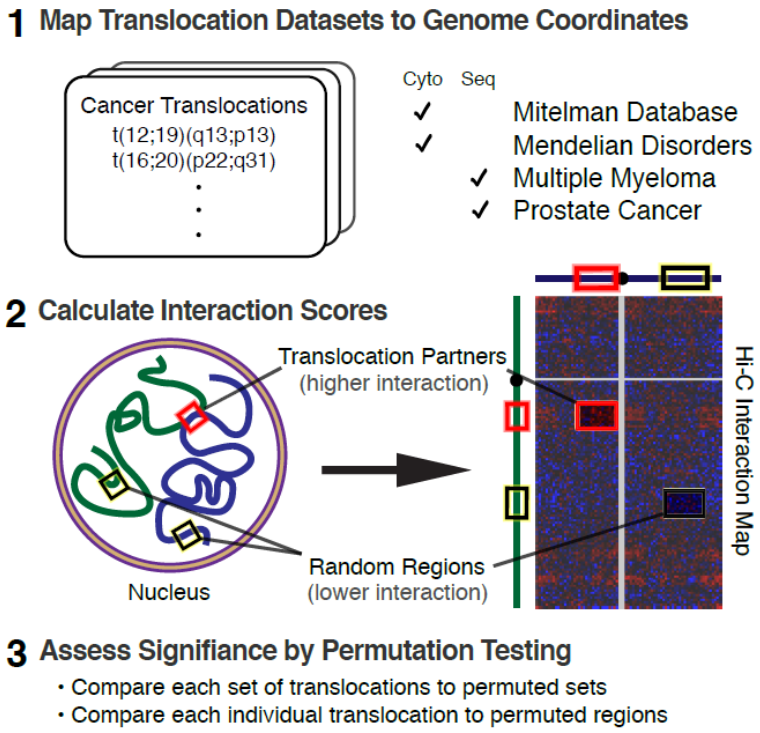
Importantly, the Mitelman Database and multiple myeloma datasets contained hundreds of translocations from lymphoid-derived malignancies, matching the cell lineage of our lymphoblastoid Hi-C data. Throughout our analysis, we focused on *inter*-chromosomal (*trans*) as opposed to *intra*-chromosomal (*cis*) events: *trans* events provide a more compelling test of the relationship between spatial proximity and partner selection [32] because loci on different chromosomes are not physically connected by a continuous DNA sequence.

Table A3.1: Translocation datasets and permutation results

Dataset	Mean Hi-C Score (Translocations)	Mean Hi-C Score (Permutations)	Permutation P-Value	Rank-Sum P-Value	Total Unique Translocations	% Genome Covered	% Interactions Covered*	Source
Mitelman Database (all)	-1.09	-1.17	<0.001	1.18E-11	577	79.7%	1.12%	Ref [31]
Mitelman Database (blood)	-1.08	-1.16	<0.001	5.67E-11	440	72.4%	0.87%	Ref [31]
Mitelman Database (non-blood)	-1.12	-1.17	<0.001	0.026	137	45.2%	0.27%	Ref [31]
Multiple Myeloma	-1.05	-1.17	<0.001	0.021	89	22.9%	0.03%	Ref [30]
Prostate Cancer	-1.09	-1.15	0.012	0.069	89	21.9%	0.03%	Ref [29]
Mendelian	-1.12	-1.16	<0.001	0.069	779	91.5%	0.79%	DACRO

*Percentage of inter-chromosomal 1-Mb bins that are covered by translocations.

To measure the level of spatial proximity between translocation partners, we devised a permutation-based approach (**Supp. Figure A3.1**). For each translocation, we mapped the chromosomal bands to genome coordinates and assigned a Hi-C interaction score by calculating the mean of the normalized read counts of all overlapping megabase bins. To assess the significance of translocation-partner interactions, we generated null distributions of Hi-C scores using five permutation methods (**Methods**). This robust permutation strategy corrected for potential biases including 1) systematic differences in association between pairs of chromosomes, 2) regions of the genome that interact with many other regions, 3) sizes and positional biases of regions in our translocation sets, and 4) broad chromatin features. Throughout, we observed similar results for all five permutation methods; we present results from Permutation Method 1 in the main text, and the others in the supplement.



Supp. Figure A3.1: Schematic diagram of our permutation approach

We applied our permutation method to test each translocation dataset for evidence of increased interaction between translocation partners. In Hi-C interaction maps from karyotypically-normal lymphoblastoid cells, translocation partners interacted more strongly than permuted regions with similar characteristics (**Table A3.1**, **Supp. Table A3.1**). Although the magnitude of the effect was incremental compared to the overall distribution of *trans* Hi-C scores (**Supp. Figure A3.2**), the finding was statistically significant ($P < 0.05$, permutation test) in all four datasets, including translocations found in multiple cancer types as well as rare Mendelian disorders (**Table A3.1**, **Figure A3.2A,C**). A closer examination of the distribution of interaction scores for true and permuted translocations showed that this signal arose from the sum of small effects across a broad range of translocations, rather than from large effects from a smaller number of rearrangements (**Figure A3.2B, A3.2D**, **Supp. Figure A3.3**), particularly for translocations from the Mitelman Database.

Permutation testing also highlighted multiple individual translocation partners that showed significant interactions (**Table A3.2**). Many of the high-scoring translocations included 19p13, the site of the *E2A* gene, which is frequently dysregulated in acute lymphoblastic leukemia (**Figure A3.2E**) [33]. Several translocations included 14q32, the chromosomal band including the *IGH* locus, reflecting the high number of double-stranded breaks in this region. For example, t(4;14)(p16;q32), is frequently found in multiple myeloma patients, causing dysregulated expression of *FGFR3* and/or *MMSET* (**Figure A3.2F**) [34, 35]. Spatial proximity may play a particularly important role in the genesis of these individual translocations.

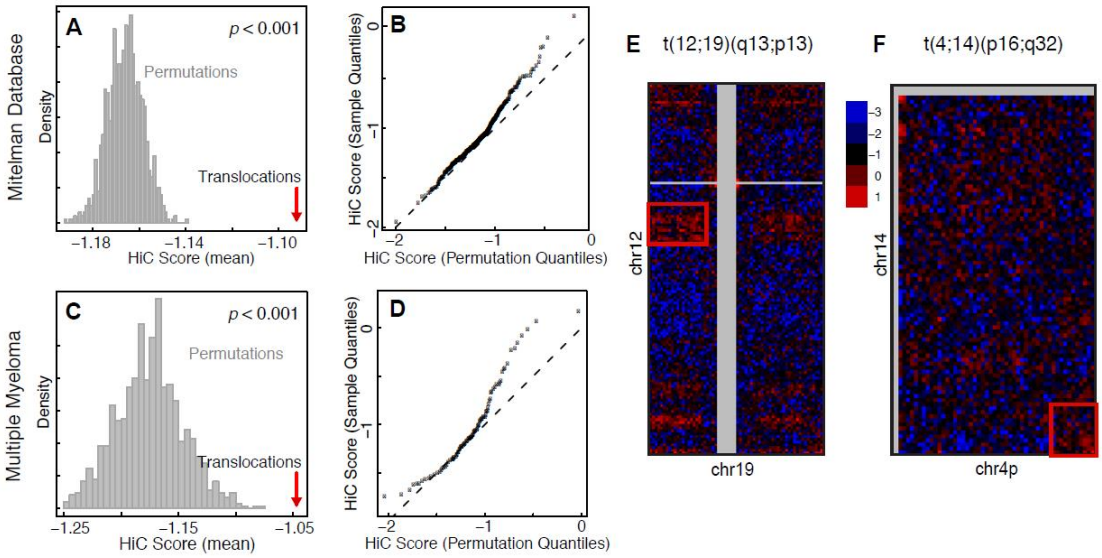


Figure A3.2*:* Many translocation partners co-localize in the normal nucleus.

Permutation test results for blood translocations from the Mitelman Database **(A and B)** and multiple myeloma whole genome sequences **(C and D)**. Histograms represent the mean Hi-C scores for 1,000 permuted sets of translocations that preserve the characteristics of the true set (Permutation Method 1, see Methods). Q-Q plots compare the permuted scores for individual translocations with the observed scores for each set. Heatmaps show Hi-C interactions in GM06990 cells for **(E)** t(4;14)(p16;q32) and **(F)** t(12;19)(q13;q13). Red boxes indicate the chromosomal bands containing the translocation breakpoints.

Tissue-specific effects. Multiple lines of evidence suggest that genome organization is highly context-dependent. Gene-level or chromosomal contacts exhibit specific and reproducible changes across tissue types and time points [27, 36, 37, 38], or in response to perturbation [5, 39]. We therefore hypothesized that the evidence for spatial proximity would be highest for translocation partners observed in malignancies derived from cells similar to a lymphoblastoid cell line. Indeed,

recurrent translocations observed in blood cancers in the Mitelman database overlapped significantly with regions of the genome that co-localize in lymphoblastoid cells ($P < 10^{-11}$, Wilcoxon rank-sum test), while the translocations observed in non-blood tumors overlapped less strongly ($P < 0.05$, Wilcoxon rank-sum test, **Table A3.1**). For example, the translocations partners for t(12;15)(p13;q15), a rearrangement found in acute lymphoblastic leukemia and lymphoblastic lymphoma, interacted much more significantly ($P < 0.001$, permutation test) in GM06690 cells than another pair of translocating loci on the same chromosomes, t(12;15)(p13;q26), found in fibrosarcoma ($P > 0.3$, permutation test). Furthermore, the set of translocations identified through sequencing of multiple myeloma ($P < 0.001$, permutation test) interacted more significantly than the set of translocations identified in prostate cancer ($P < 0.012$, permutation test). These results suggest that tissue-specific changes in genome organization may predispose specific regions to translocate in different malignancies.

Table A3.2: Previously known translocation-prone loci that significantly co-localize in normal nuclei

Dataset	Karyotype	Permutation P-Value	Disease
Mitelman Database	t(12;19)(q13;p13)	<0.001	Multiple myeloma
	t(16;20)(p13;q13)	<0.001	Multiple myeloma
	t(12;15)(p13;q15)	<0.001	Acute lymphoblastic leukemia/lymphoblastic lymphoma
	t(1;19)(q22;p13)	<0.001	Acute lymphoblastic leukemia/lymphoblastic lymphoma
	t(3;19)(p21;p13)	<0.001	Acute lymphoblastic leukemia/lymphoblastic lymphoma
	tas(19;22)(q13;q13)	<0.001	Giant cell tumor of bone
Multiple Myeloma	t(11;X)(p11.12;p11.1)	<0.001	Multiple myeloma
	t(4;14)(p16.3;q32.33)	0.049	Multiple myeloma
Prostate Cancer	none	-	-
Mendelian Disease	t(16;20)(p13.3;q13.33)	<0.001	Developmental delay, polycystic kidneys, ventricular septal defect, pulmonary stenosis
	t(12;14)(q24;q32)	<0.001	Coffin-Siris syndrome
	t(4;8)(p16;p23)	<0.001	Waardenburg syndrome, type IIC
	t(X;4)(q21;q13)	<0.001	Ovarian dysgenesis and primary amenorrhoea

Chromatin status of translocation partners. One explanation for the observed frequency of interaction between translocation partners is that these regions have preferentially high gene content, or lie in the open chromatin compartment, rendering them easily accessible and mutable. Alternatively, proximity-mediated rearrangements might occur in both the open and closed chromatin compartments. To distinguish between these models, we repeated our permutation method, scoring genomic regions for gene content and chromatin compartment score rather than on interaction

frequency. We found that in all four datasets, translocation breakpoints were significantly enriched for gene-rich, euchromatic regions (**Figure A3.3A-B**).

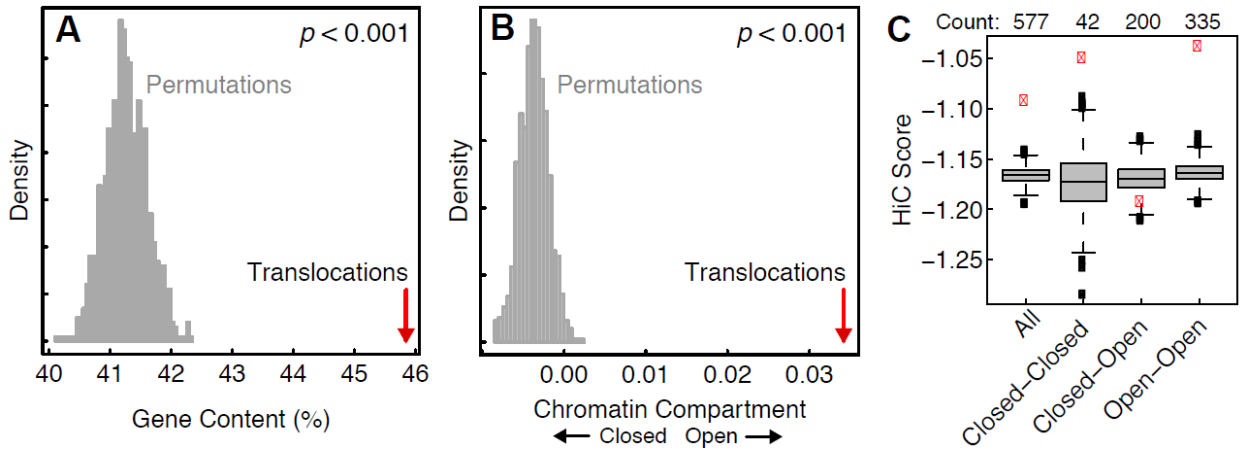


Figure A3.3: Features of translocation breakpoints

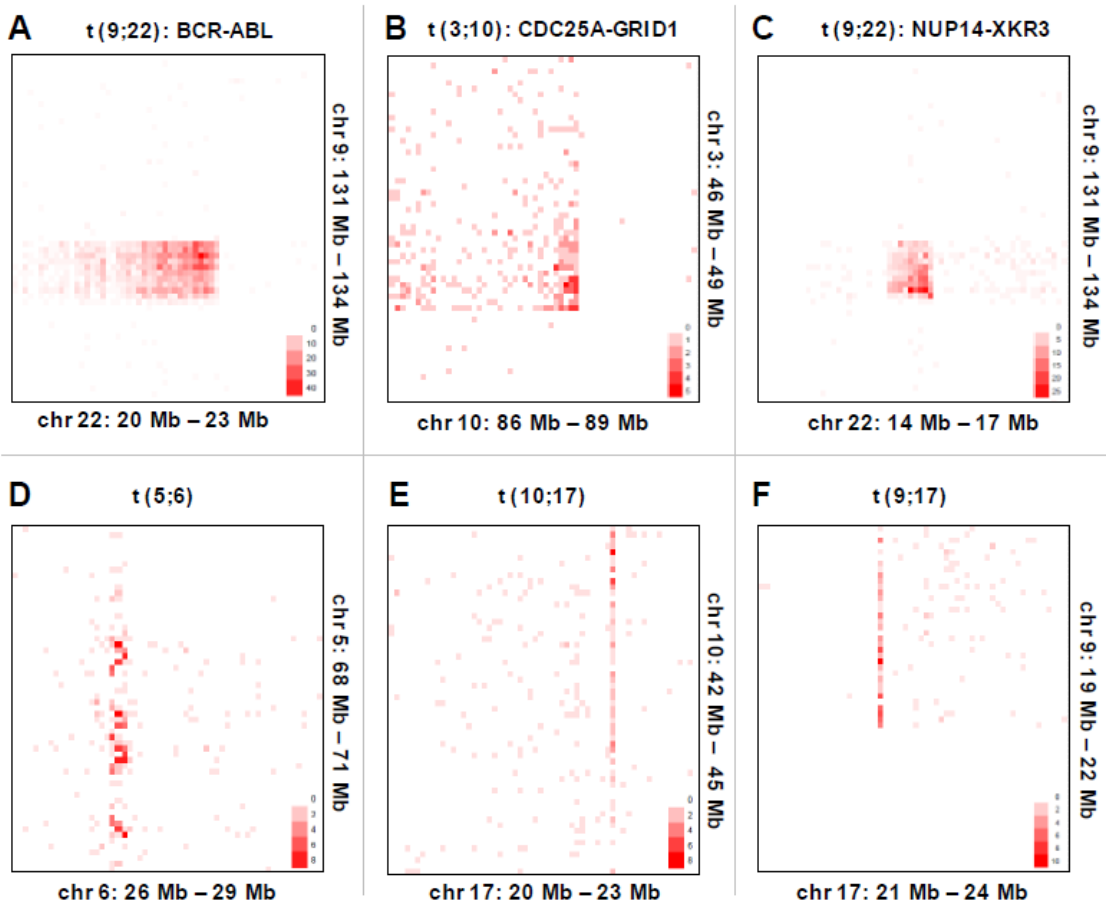
Comparison of average **(A)** gene content (% bases spanned by transcripts; includes both exons and introns) and **(B)** chromatin compartment score for translocations (red) and 1,000 permuted sets that preserve the characteristics of the true set. Chromatin compartment scores are calculated using the first eigenvector of the *inter*-chromosomal Hi-C data (see Methods). Positive and negative scores indicate open and closed chromatin compartments, respectively. **(C)** Mean Hi-C interaction scores for Mitelman blood translocations (red) compared to sets of permuted regions selected from the same chromatin compartments (gray).

To control for the contribution of chromatin state to translocation partner interactions, we repeated our permutation tests controlling for chromatin compartment (open versus closed, see Methods). We found that even after controlling for chromatin compartment, translocation partners interacted more strongly than expected by chance, although the significance of this finding was reduced across all datasets (**Figure A3.3C**, **Supplemental Tables A3.1-2**). Translocations whose partners resided in the same chromatin compartment interacted more strongly and more significantly than translocations with one partner in each compartment (**Figure A3.3C**). While translocation partners are more likely to reside in euchromatic regions, spatial proximity within the nucleus provides additional information.

Fine-mapping translocation breakpoints in the K562 cell line. The Hi-C data provided evidence not only for pre-translocation chromosomal proximity in both normal and K562 cell lines, but also for translocations that have already occurred. As seen in the case of the *BCR-ABL*

translocation (**Figure A3.2A**), the signal strength for existing translocations is much higher than subtler signal of pre-translocation interactions. We thus attempted to use Hi-C data to fine-map chromosomal breakpoints in K562 cells. Although K562 is a widely used cancer cell, the exact breakpoints and genes implicated its translocations remain largely unknown. We assembled a set of 13 previously described, coarsely-mapped translocations in K562 cells (**Table A3.3**), and explored the raw Hi-C data at these loci in order to locally refine the intervals containing the DNA breakpoints. We observed a local hotspot of inter-chromosomal reads for 6 of these 13 previously described rearrangements. That we did not observe evidence for half of the previously reported translocations is not entirely surprising: K562 lines in culture across different laboratories may have divergent karyotypes, though all share the *BCR-ABL* driver translocation. Moreover, the reported cytogenetic rearrangements were not validated by secondary methods.

At the 6 loci at which Hi-C signal was detected, we counted the number of raw Hi-C reads mapping to 50-kilobase intervals in the broad region of peak interaction signal. At 3 of the 6 regions, a heat map showing read counts at 50-kb resolution showed a pattern of peak signal at a corner, with signal decaying in a single direction along both chromosomes (**Figure A3.4**, **Supp. Figure A3.3A-C**). This is an expected pattern for an unbalanced translocation; signal is highest at the location of the chromosomal breakpoints, where ligation has occurred, and decays with distance away from the breakpoints along the fused chromosomal bands. For the *BCR-ABL* translocation, for example, we mapped the breakpoint to a 50-kb region spanning chr9:132,550,000-132,600,000 and chr22:21,950,000-22,000,000 (**Figure A3.4A**, **Table A3.3**). Within this region, three reads (the highest local density) mapped to within a kilobase of the described precise breakpoint for the *BCR-ABL* translocation in the K562 cell line (chr9:132,596,950-132,597,013 and chr22:21,962,697-21,962,754) [40].



Supp. Figure A3.3: Fine mapping of previously reported inter-chromosomal translocations in K562.

Heat maps above show the observed number of reads mapping to 50kb x 50kb bins at the BCR-ABL locus (A), the CDC25A-GRID1 locus (B), the NUP214-XKR3 locus (C), as well as three other loci sites of previously reported inter-chromosomal translocations in the K562 cell line (D, E, F). At coarse resolution (1Mb x 1Mb, as seen in the Hi-C Data Browser), all six of these loci showed patterns characteristics of unbalanced translocation, but at fine 50kb resolution only 3 (those shown in A, B, and C) show a pattern of peak signal in a corner with signal decaying along each chromosome in one direction. In A-C, the 50kb bin with peak signal is most likely to contain the translocation

We next attempted to identify the breakpoint region for the other loci showing Hi-C signal characteristic for a translocation in the K562 cell line (**Table A3.3**, **Supp. Figure A3.3B-C**). Sequence-based identification of fusion gene transcripts in the K562 cell line has recently identified a second translocation between chr9 and chr22 involving the gene partners NUP214 and XKR3 [41]. The Hi-C data also showed clear evidence for a translocation between these loci; fine-mapping at the 50-kilobase scale revealed a likely breakpoint at chr9:133,050,000-133,100,000 and

chr22:15,650,000-15,700,000. Further fine-mapping revealed that 14 reads (nearly all reads mapping to the 50-kilobase region) clustered tightly within a region at chr9:133,064,000-133,065,000 and chr22:15,680,000-15,681,000, suggesting that the breakpoint lies between the HindIII restriction sites in these regions (see Methods).

Table A3.3: Fine-mapping known K652 translocation breakpoints in Hi-C data

Source	Reported K562 translocation	Evidence in Hi-C data?	Breakpoint region (hg18)	Annotated gene closest to predicted breakpoint
K562 Karyotype [42]	t(9;22)	Yes	chr9:132,550,000–132,600,000	chr9: ABL1
			chr22:21,950,000–22,000,000	chr22: BCR
	t(3;10)	Yes	chr3:48,150,000 – 48,200,000	**chr3: CDC25A (cell cycle division 25A isoform A)**
			chr10:87,800,000 – 87,850,000	chr10: GRID1
	t(10;17)	Yes	chr10:42,200,000–42,250,000	**chr10: BC039000 (Homo sapiens cyclin Y-like 2)**
			chr17:22,150,000–22,200,000	chr17: **enhancer marks** in K562
	t(9;17)	Yes	chr9:20,150,000–20,200,000	**chr9: MLLT3 (myeloid/lymphoid or mixed-lineage leukemia, translocated to 3)**
			chr17:22,150,000–22,200,000	chr17: **enhancer marks** in K562
	t(5;6)	Yes	chr5:69,000,000–69,050,000	chr5: GUSBP3
			chr6:27,000,000–27,050,000	chr6: GUSBP1
	t(9;13)	No	-	-
	t(1;21)	No	-	-
	t(2;19)	No	-	-
	t(19;20)	No	-	-
	t(6;11)	No	-	-
	t(12;19)	No	-	-
K562 Next-Gen Sequencing [40, 41]*	t(9;22)_2	Yes	chr9: 133,050,000–133,100,000	chr9: **NUP214**
			chr22: 15,650,000–15,700,000	chr22: XKR3
	t(1;11)	No	-	-

*Targeted sequencing of 476 cancer-related gene cDNA transcripts only; did not detect non-coding breakpoints or translocations affecting other genes

Finally, we applied a similar fine-mapping procedure to identify breakpoint regions for a translocation t(3;10) that had been previously reported in a cytogenetic study [42]. This rearrangement has previously only been described based on its visual karyotypic appearance, and even the cytogenetic bands involved in this translocation have yet not been identified. The Hi-C data showed clear evidence for a translocation (**Figure A3.4B**); we were able to fine-map the breakpoint, for the first time, to a region spanning chr3:48,150,000-48,200,000 and chr10:87,800,000-87,850,000. Hi-C read data was too sparse to perform further fine-mapping, but these regions

overlap the CDC25A (cell cycle division 25A isoform A) transcript on chr3 and the GRID1 transcript on chr10.

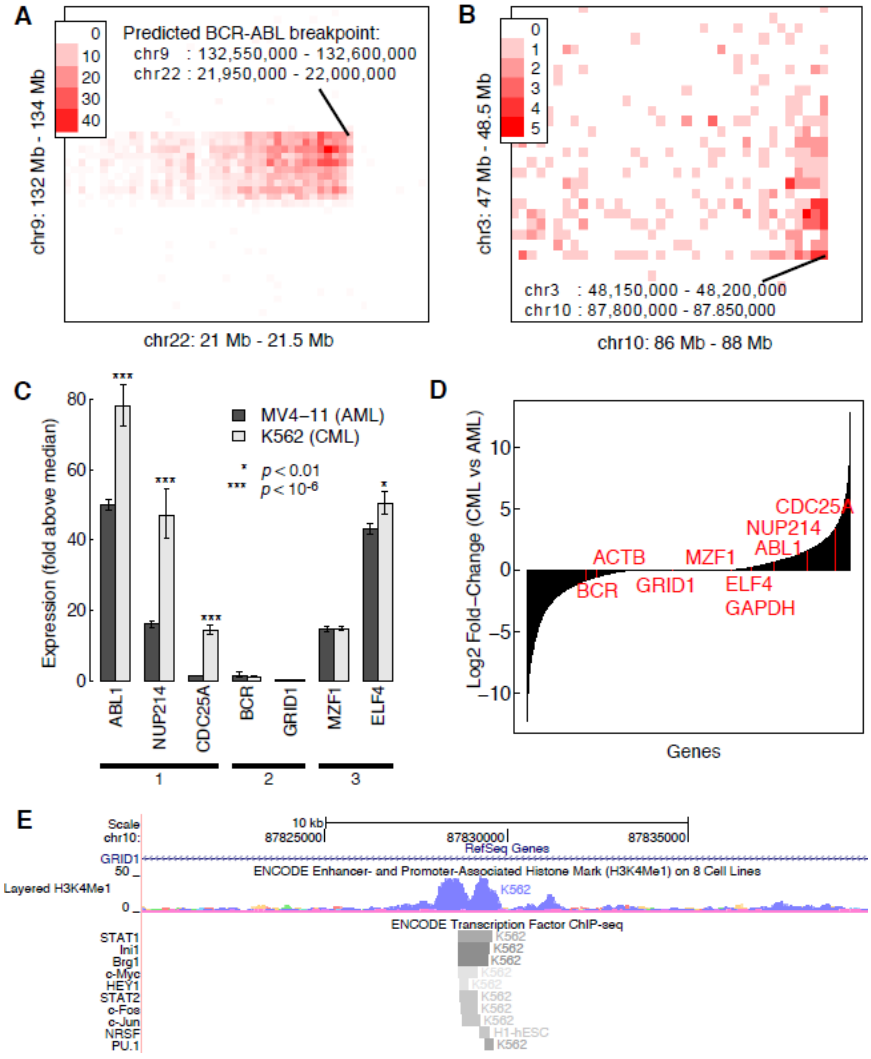


Figure A3.4: Fine mapping of previously reported translocations in the K562 cell line.

Heat maps showing the observed number of reads mapping to 50Kb bins at selected regions of **(A)** BCR-ABL and **(B)** the novel t(3;10) CDC25A-GRID1 translocation. **(C)** Gene expression for dysregulated translocation partners (1), normally-regulated translocation partners (2), and constitutively-expressed myeloid genes (3) in MV4-11 (AML) and K562 (CML) cell lines. Expression values for each gene are normalized to the median expression for all genes. Note that XKR3, the translocation partner of NUP214, is not assayed on this microarray platform. **(D)** CML-to-AML $\log_2$ fold-change for all assayed genes, sorted in increasing order. Red lines indicate the fold-change for labeled genes. The dysregulated translocation partners CDC25A, NUP214, and ABL1 are highly up-regulated in CML, all falling in the upper quartile of genes in terms of fold-change. **(E)** ENCODE ChIP-seq data for transcription factors and H3K4me1 near the predicted GRID1 breakpoint. Color for H3K4me1 corresponds to cell type (only K562 shows significant signal in this region). Color for transcription factor data is proportional to the ChIP-seq signal. Data was viewed with the UCSC Genome Browser, genome build *hg18* (http://genome.ucsc.edu) [50].

Since CDC25A is an oncogene required for progression from G1 to S phase, we hypothesized that this translocation may result in aberrantly upregulated CDC25A. When we compared the expression of CDC25A in K562 cells [43] and an AML cell line [44], we found that CDC25A was approximately 8-fold higher in K562 cells ($P < 10^{-6}$, Student's *t* test). The significance and magnitude of this result was comparable to other dysregulated translocation partners in K562 cells (ABL1 and NUP214, **Figure A3.4C-D**). Translocation partners that supply regulatory elements (BCR, GRID1) and genes constitutively expressed in the myeloid lineage (MZF1, ELF4, and GAPDH) were not significantly upregulated. Interestingly, the breakpoint region on chr10 maps to an intronic region of the GRID1 transcript; within this 50-kilobase region, there are strong K562-specific H3K4Me1 histone marks and ChIP-seq derived transcription factor binding sites that are not present in any other ENCODE cell type, suggesting that perhaps the fusion event either brings CDC25A under control of an existing enhancer, or creates a novel regulatory element that might drive CDC25A expression (**Figure A3.4D**). Our results suggest that the GRID1-CDC25A translocation may represent a novel functional fusion, although further characterization of this rearrangement will be required to define its exact functional role.

Discussion

Given the frequency of driver genomic rearrangements in multiple tumor types, identification of the specific genomic loci as well as the predisposing factors for transformative translocations have important implications for the genesis of cancer. Here we provide evidence that many translocation breakpoints spatially co-localize in the normal genome, suggesting a broad role for proximity in determining the frequency of translocations between partner loci. Four complementary collections of rearrangements yielded concordant results, albeit of varying significance, suggesting that our conclusions are robust to the differences in bias and selection present in these datasets. Our results also are consistent after controlling for differences in chromosome positioning, region size, mappability, coverage, and chromatin state, although it is conceivable that additional, unidentified variables are producing a synthetic association.

Several factors may explain the quantitatively small, albeit statistically significant, increase in Hi-C interactions for translocation partners. First, interactions between translocation-prone loci may be transient, or occur in only a small fraction of the assayed cells. Second, it is theoretically possible that subpopulations of the cultured cells in fact harbor *de novo* translocations that we detect as a weak signal; indeed, expression of the *BCR-ABL* fusion gene can be detected at a very low level in the blood of nearly a third of healthy adults [45]. We note, however, that the distribution of *trans* reads across the chromosome cannot be entirely explained by this phenomenon.

Notably, our analysis identified a difference in significance between translocations from hematologic and non-hematologic cancers, suggesting that the relationship between translocation frequency and spatial interaction is tissue-specific. At the same time, datasets of translocation from other tissues still passed the threshold for significance. This suggests that the conformational variation of the genome may not vary drastically between cell types, and that some translocation partners may interact regardless of cell type. Additional Hi-C experiments in multiple matched cell types may help to elucidate lineage-dependent variation in global chromosomal conformation and its contribution to translocation partner selection.

Recurrent translocations in the Mitelman Database produced an interaction signal substantially higher than that of the primary translocation datasets. We speculate that this is due not only to the larger sample size of the Mitelman dataset, but also to the higher proportion of stochastic, passenger rearrangements present in the primary cancer datasets. Surprisingly, the set of prostate tumor translocations interacted more significantly than the set of multiple myeloma translocations. This finding may be consistent with recent observations that some translocations in prostate cancer may occur through specific proximity-mediated mechanisms [6]. It is possible that there is a systematic mechanistic difference in translocation genesis that is unique to prostate cancer, especially given the relatively high rate of observed rearrangements in these genomes.

Our results support a model where translocation partners reside in broad interacting domains that span multi-megabase chromosomal regions. Indeed, when we examined the 1-Mb region

surrounding each breakpoint in the primary tumor datasets, we did not observe significantly elevated local signal (**Supp. Table A3.1**). This suggests that it is not just the breakpoint loci that contact each other prior to translocation, but also the larger chromatin domains that contain them. These broad interactions, occurring across an aggregate cell population, bring translocation partners into close spatial proximity, increasing the likelihood of rearrangement.

While we report that chromosome conformation correlates with translocation partner selection, consistent with the "contact first" model of genomic rearrangements, some translocations may arise in part or entirely by other mechanisms. In particular, the "breakage first" hypothesis suggests that the ends of double-stranded breaks are highly mobile, and that cellular mechanisms exist to gather such lesions after a breakage has occurred, potentially over a large distance [46]. It is possible that these mechanisms are responsible for the chromosomal rearrangements that did not show elevated Hi-C signal in this study. Context-specific induction of proximity could also explain a subset of these results. To dissect the contributions of spatial proximity, double-stranded break mobility, and other cellular processes to translocation partner selection, investigators will need to examine translocations induced by experimental mutagenesis, prior to selection in the tumor microenvironment.

Finally, our results demonstrate the utility of the genome-wide chromosome conformation capture approach in mapping existing translocation breakpoints to kilobase resolution. Other groups have recently demonstrated targeted and genome-wide sequencing-based methods to resolve translocation breakpoints [29, 30, 40, 41]. However, we suggest that Hi-C may provide an alternative and more sensitive method for detecting translocations genome-wide compared to other methods, since genomic rearrangements produce robust signals involving regions up to a megabase from the breakpoints. These long-range interactions may allow for sequence-based karyotyping by illuminating the linkage between multiple breakpoints on rearranged chromosomes.

Given a role for spatial proximity in translocation partner selection, the molecular mechanisms that govern three-dimensional genomic architecture in normal and cancerous cells may

prove etiologically important in our understanding of oncogenic transformation. Work to characterize the interactions between chromosome conformation and triggers for rearrangements may provide a foundation for advances in prevention and treatment of these malignancies.

Methods

Translocation datasets. We collected four large inter-chromosomal translocation datasets, derived both from karyotyping and high-throughput sequencing studies. First, we collected a set of recurrent *trans*-chromosomal cancer translocations that have been observed in multiple patient cases from the Mitelman Database, available at the NCI Cancer Genome Anatomy Project website (http://cgap.nci.nih.gov/Info/CGAPDownload/). The Mitelman Database describes translocations using chromosomal bands; precise breakpoints were not available. The average size of defined chromosomal translocation bands in this database was large (~10Mb). While positive selection modifies the frequency of cancer translocations, particularly driver rearrangements, we expected that many of these recurrent translocations were predisposed to recur due to factors such as genome organization.

Translocations from multiple myeloma and prostate cancer were identified from whole-genome or exome sequences using the *dRanger* algorithm (Drier Y. *et al.,* submitted). We used translocations with at least three supporting reads in our analysis. Multiple myeloma translocations were obtained from Chapman et al. [30]. Prostate cancer translocations were obtained from the supplement of Berger et al. [29]. Compared to the Mitelman Database, we expected catalogs of translocations in primary tumors to contain a higher frequency of passenger rearrangements, as well as a higher proportion of private mutations that occurred stochastically rather than systematically due to predisposing factors.

Finally, we collected all two-partner inter-chromosomal translocations (n=947) associated with Mendelian syndromes from the Disease Associated Chromosomal Rearrangement Database (https://www1.hgu.mrc.ac.uk/Softdata/Translocation/). Again, precise breakpoints were not available

for this dataset. Because these translocations can cause severe phenotypes, many of these mutations are not transmitted through generations: these diseases, though rare, are caused by recurrent *de novo* translocations. In addition, these genomic rearrangements do not experience the same positive selective pressures as the cancer translocations, complementing the previous datasets.

For the first Mitelman and Mendelian translocation databases, we mapped the cytogenetic bands (e.g. t(9;22)(p13;q13)) to human genome coordinates using the UCSC Genome Browser Build hg18. We considered 3-way translocations as 3 distinct two-way translocations, and excluded all translocations involving more than 3 partners. We also excluded translocations involving entire chromosomal arms, and did not include any inter-chromosomal rearrangements (e.g. inversions).

Chromosomal conformation capture. We used public Hi-C data (GEO accession GSE18199) generated to interrogate the long-range genomic interactions in the GM06690 lymphoblastoid and K562 erythroleukemic cell lines [18]. We used the processed, mapped reads and the one-megabase binning scheme as described. To control for differences in coverage, number of HindIII sites, mappability, and other features unique to each one-megabase bin, we normalized read counts within each bin using the universe of all inter-chromosomal (*trans*) reads: *Hi-C Score* = log_2 [(number of *trans*-reads with one mate-pair mapping to region 1) × (number of *trans*-reads with one mate-pair mapping to region 2) / (total number of *trans*-reads in entire dataset)]. We applied the *log* variance-stabilizing transformation to reduce the contributions of strong outliers when calculating summary statistics over a region.

Chromatin compartment and gene content. We assigned regions to chromatin compartments using principal component analysis as described [18]. Positive and negative scores indicate open and closed chromatin compartments, respectively, and correlate with other genomic features such as gene content, histone modification, and DNAseI hypersensitivity. For each translocation region, we calculated a compartment score as the mean of the principal component

values for all overlapping megabase bins. We represented gene content as the percentage of bases covered by RefSeq genes, including both exons and introns.

Permutation testing. We employed a permutation strategy to search Hi-C data for evidence of interaction between translocation breakpoints. We calculated the interaction between each translocation region as the mean Hi-C score of all overlapping megabase bins. We obtained similar results when considering the median instead of the mean (data not shown). When calculating these summary statistics, we did not include bins that 1) overlapped centromeres or 2) had no coverage across the entire dataset. To assess the significance of individual translocations, we generated a null distribution by considering 1,000 random pairs of regions with one of four permutation methods:

1) We selected regions of identical size from the same chromosome pair. This within-chromosome permutation scheme controlled for the systematic differences in association between pairs of chromosomes: smaller gene-rich chromosomes, for instance, tend to group together [15].
2) We fixed one region, and selected as a partner a random region of identical size on the same chromosome. This controlled for features of the translocation partners that might predispose them to interact with many other regions on the same chromosome.
3) We fixed one region, and selected as a partner a random region of identical size on any other chromosome. This controlled for features of the translocation partners that might predispose them to interact with many other regions across the genome.
4) We fixed one region, and selected as a partner a random region from the entire set of translocations partners that did not fall on the same chromosome as the fixed partner.

In all cases where we selected random regions, we required that less than 50% of the bins in the random region overlapped with centromeric regions or bins with no coverage across the entire dataset.

For each individual translocation, we calculated the *p*-value for each translocation as the fraction of permuted locations that exceeded its interaction score, and corrected for multiple hypothesis testing using the Benjamini-Hochberg method.

We assessed the significance of each translocation dataset as a group using a similar approach. For each of our four datasets, we generated 1,000 randomized datasets that preserved

the overall properties of the group of translocations: the chromosome pairings and region sizes matched the original set. We calculated a summary score for each of these randomized datasets that represented the mean interaction score across all translocations, and calculated a *p*-value by comparing these statistics to the null distributions. We also assessed the differences between the interaction scores of the true and randomized data distributions using the Student's *t* test and Wilcoxon rank sum test to monitor the degree to which outliers drove the result in the permutation scheme.

Permutations within chromatin compartment. We also evaluated the significance of our results by controlling for chromatin compartment in Permutation Methods 1-3. To accomplish this, we allowed swapping only within compartments. For each translocation partner, we calculated the chromatin score and chosen randomly from similarly-sized regions whose chromatin scores had the same sign.

Fine-mapping of translocation breakpoints. To identify likely chromosomal breakpoints responsible for previously reported translocations, we first identified the 1Mb x 1Mb bin across the chromosome pair with the highest total normalized read count. We then selected all reads mapping to a 3Mb x 3Mb window around this bin. We then counted the number of observed reads mapping to 50Kb x 50Kb bins, and looked for a pattern characteristic of unbalanced translocation. We then selected the corner-most 50Kb bin, and counted reads mapping to 1Kb x 1Kb regions within this larger bin. In some cases, the read count was sufficient to allow breakpoint identification at this fine scale, but in other cases read coverage was too sparse to further localize the breakpoint. In all cases, resolution is limited by the density of HindIII restriction sites (*i.e.*, the sites at which DNA is cleaved during the Hi-C experiment). Heat maps of raw read count at the 50-kilobase pair scale are shown in supplementary materials.

Gene expression analysis. We downloaded publicly available microarray data from the NCBI Gene Expression Omnibus for K562 (CML) and MV4-11 (AML) cell lines from GSE12056 [43] and GSE26114 [44], respectively. Both experiments used the Affymetrix HG-U133 Plus 2.0 array

platform. We preprocessed and normalized the raw CEL files using the GC-RMA algorithm [47] implemented in R Bioconductor [48], and mapped probes to genes using the custom CDF from BrainArray [49]. Although the distributions of gene expression appeared similar, we quantile-normalized the data to ensure that arrays from the two different cell lines were comparable. In Figure A3.4C, we calculated the fold-change between each gene and the median expression for all genes. In Figure A3.4D, we calculated the fold-change between the average values of three replicate (for MV4-11) and nine replicate (for K562) arrays.

Acknowledgements

The authors thank Geoffrey Fudenberg, Rachel McCord, Job Dekker, and Levi Garraway for discussions and critiques.

References

1. Larson RA, Golomb HM, Rowley JD. (1981) Chromosome changes in hematologic malignancies. CA Cancer J Clin 31: 222-38.
2. Greaves MF, Wiemels J. (2003) Origins of chromosome translocations in childhood leukaemia. Nat Rev Cancer 3: 639-49.
3. Brenner JC, Chinnaiyan AM. (2009) Translocations in epithelial cancers. Biochim Biophys Acta 1796: 201-15.
4. Edwards PAW. (2010) Fusion genes and chromosome translocations in the common epithelial cancers. J Pathol 220: 244-54.
5. Mani R, Tomlins SA, Callahan K, Ghosh A, Nyati MK, Varambally S, Palanisamy N, Chinnaiyan AM. (2009) Induced chromosomal proximity and gene fusions in prostate cancer. Science 326: 1230.
6. Haffner MC, Aryee MJ, Toubaji A, Esopi DM, Albadine R, Gurel B, Isaacs WB, Bova GS, Liu W, Xu J, Meeker AK, Netto G, De Marzo AM, Nelson WG, Yegnasubramanian S. (2010) Androgen-induced TOP2B-mediated double-strand breaks and prostate cancer gene rearrangements. Nat Genet 42: 668-75.
7. Nowell P, Hungerford D. (1960) A minute chromosome in human chronic granulocytic leukemia. Science 132
8. Rowley JD. (1980) Chromosome abnormalities in human leukemia. Annu Rev Genet 14: 17-39.
9. Koeffler HP, Golde DW. (1981) Chronic myelogenous leukemia--new concepts (first of two parts). N Engl J Med 304: 1201-9.
10. Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, Sun X, Varambally S, Cao X, Tchinda J, Kuefer R, Lee C, Montie JE, Shah RB, Pienta KJ, Rubin MA, Chinnaiyan AM. (2005) Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science 310: 644-8.
11. De S, Michor F. (2011) DNA secondary structures and epigenetic determinants of cancer genome evolution. Nat Struct Mol Biol 18: 950-5.
12. Kozubek S, Lukásová E, Marecková A, Skalníková M, Kozubek M, Bártová E, Kroha V, Krahulcová E, Slotová J. (1999) The topological organization of chromosomes 9 and 22 in cell nuclei has a determinative role in the induction of t(9,22) translocations and in the pathogenesis of t(9,22) leukemias. Chromosoma 108: 426-35.
13. Meaburn KJ, Misteli T, Soutoglou E. (2007) Spatial genome organization in the formation of chromosomal translocations. Semin Cancer Biol 17: 80-90.
14. Parada L, Misteli T. (2002) Chromosome positioning in the interphase nucleus. Trends Cell Biol 12: 425-32.
15. Bolzer A, Kreth G, Solovei I, Koehler D, Saracoglu K, Fauth C, Müller S, Eils R, Cremer C, Speicher MR, Cremer T. (2005) Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. PLoS Biol 3: e157.

16. Cremer T, Cremer M, Dietzel S, Müller S, Solovei I, Fakan S. (2006) Chromosome territories--a functional nuclear landscape. Curr Opin Cell Biol 18: 307-16.
17. Fraser P, Bickmore W. (2007) Nuclear organization of the genome and the potential for gene regulation. Nature 447: 413-7.
18. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J. (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326: 289-93.
19. Lukásová E, Kozubek S, Kozubek M, Kjeronská J, Rýznar L, Horáková J, Krahulcová E, Horneck G. (1997) Localisation and distance between ABL and BCR genes in interphase nuclei of bone marrow cells of control donors and patients with chronic myeloid leukaemia. Hum Genet 100: 525-35.
20. Roix JJ, McQueen PG, Munson PJ, Parada LA, Misteli T. (2003) Spatial proximity of translocation-prone gene loci in human lymphomas. Nat Genet 34: 287-91.
21. Mathas S, Kreher S, Meaburn KJ, Jöhrens K, Lamprecht B, Assaf C, Sterry W, Kadin ME, Daibata M, Joos S, Hummel M, Stein H, Janz M, Anagnostopoulos I, Schrock E, Misteli T, Dörken B. (2009) Gene deregulation and spatial genome reorganization near breakpoints prior to formation of translocations in anaplastic large cell lymphoma. Proc Natl Acad Sci U S A 106: 5831-6.
22. Branco MR, Pombo A. (2006) Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. PLoS Biol 4: e138.
23. Tolhuis B, Palstra RJ, Splinter E, Grosveld F, de Laat W. (2002) Looping and interaction between hypersensitive sites in the active beta-globin locus. Mol Cell 10: 1453-65.
24. Spilianakis CG, Flavell RA. (2004) Long-range intrachromosomal interactions in the T helper type 2 cytokine locus. Nat Immunol 5: 1017-27.
25. Osborne CS, Chakalova L, Brown KE, Carter D, Horton A, Debrand E, Goyenechea B, Mitchell JA, Lopes S, Reik W, Fraser P. (2004) Active genes dynamically colocalize to shared sites of ongoing transcription. Nat Genet 36: 1065-71.
26. Dekker J, Rippe K, Dekker M, Kleckner N. (2002) Capturing chromosome conformation. Science 295: 1306-11.
27. Neves H, Ramos C, da Silva MG, Parreira A, Parreira L. (1999) The nuclear topography of ABL, BCR, PML, and RARalpha genes: evidence for gene proximity in specific phases of the cell cycle and stages of hematopoietic differentiation. Blood 93: 1197-207.
28. Hecht JL, Aster JC. (2000) Molecular biology of Burkitt's lymphoma. J Clin Oncol 18: 3707-21.
29. Berger MF, Lawrence MS, Demichelis F, Drier Y, Cibulskis K, Sivachenko AY, Sboner A, Esgueva R, Pflueger D, Sougnez C, Onofrio R, Carter SL, Park K, Habegger L, Ambrogio L, Fennell T, Parkin M, Saksena G, Voet D, Ramos AH, Pugh TJ, Wilkinson J, Fisher S, Winckler W, Mahan S, Ardlie K, Baldwin J, Simons JW, Kitabayashi N, MacDonald TY, Kantoff PW, Chin L, Gabriel SB, Gerstein MB, Golub TR, Meyerson M, Tewari A, Lander ES, Getz G, Rubin MA, Garraway LA. (2011) The genomic complexity of primary human prostate cancer. Nature 470: 214-20.
30. Chapman MA, Lawrence MS, Keats JJ, Cibulskis K, Sougnez C, Schinzel AC, Harview CL, Brunet J, Ahmann GJ, Adli M, Anderson KC, Ardlie KG, Auclair D, Baker A, Bergsagel PL, Bernstein BE, Drier Y, Fonseca R, Gabriel SB, Hofmeister CC, Jagannath S, Jakubowiak AJ, Krishnan A, Levy J, Liefeld T, Lonial S, Mahan S, Mfuko B, Monti S, Perkins LM, Onofrio R, Pugh TJ, Rajkumar SV, Ramos AH, Siegel DS, Sivachenko A, Stewart AK, Trudel S, Vij R, Voet D, Winckler W, Zimmerman T, Carpten J, Trent J, Hahn WC, Garraway LA, Meyerson M, Lander ES, Getz G, Golub TR. (2011) Initial genome sequencing and analysis of multiple myeloma. Nature 471: 467-72.
31. Mitelman F, Johannson B, Mertens FE. Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer.
32. Wijchers PJ, de Laat W. (2011) Genome organization influences partner selection for chromosomal rearrangements. Trends Genet 27: 63-71.
33. Khalidi HS, O'Donnell MR, Slovak ML, Arber DA. (1999) Adult precursor-B acute lymphoblastic leukemia with translocations involving chromosome band 19p13 is associated with poor prognosis. Cancer Genet Cytogenet 109: 58-65.
34. Chesi M, Nardini E, Brents LA, Schröck E, Ried T, Kuehl WM, Bergsagel PL. (1997) Frequent translocation t(4;14)(p16.3;q32.3) in multiple myeloma is associated with increased expression and activating mutations of fibroblast growth factor receptor 3. Nat Genet 16: 260-4.

35. Malgeri U, Baldini L, Perfetti V, Fabris S, Vignarelli MC, Colombo G, Lotti V, Compasso S, Bogni S, Lombardi L, Maiolo AT, Neri A. (2000) Detection of t(4;14)(p16.3;q32) chromosomal translocation in multiple myeloma by reverse transcription-polymerase chain reaction analysis of IGH-MMSET fusion transcripts. Cancer Res 60: 4058-61.
36. Parada LA, McQueen PG, Misteli T. (2004) Tissue-specific spatial organization of genomes. Genome Biol 5: R44.
37. Hou C, Dale R, Dean A. (2010) Cell type specificity of chromatin organization mediated by CTCF and cohesin. Proc Natl Acad Sci U S A 107: 3651-6.
38. Eskiw CH, Cope NF, Clay I, Schoenfelder S, Nagano T, Fraser P. (2011) Transcription Factories and Nuclear Organization of the Genome. Cold Spring Harb Symp Quant Biol
39. Hu Q, Kwon Y, Nunez E, Cardamone MD, Hutt KR, Ohgi KA, Garcia-Bassets I, Rose DW, Glass CK, Rosenfeld MG, Fu X. (2008) Enhancing nuclear receptor-induced transcription requires nuclear motor and LSD1-dependent gene networking in interchromatin granules. Proc Natl Acad Sci U S A 105: 19199-204.
40. Ge H, Liu K, Juan T, Fang F, Newman M, Hoeck W. (2011) FusionMap: detecting fusion genes from next-generation sequencing data at base-pair resolution. Bioinformatics
41. Levin JZ, Berger MF, Adiconis X, Rogov P, Melnikov A, Fennell T, Nusbaum C, Garraway LA, Gnirke A. (2009) Targeted next-generation sequencing of a cancer transcriptome enhances detection of sequence variants and novel fusion transcripts. Genome Biol 10: R115.
42. Naumann S, Reutzel D, Speicher M, Decker HJ. (2001) Complete karyotype characterization of the K562 cell line by combined application of G-banding, multiplex-fluorescence in situ hybridization, fluorescence in situ hybridization, and comparative genomic hybridization. Leuk Res 25: 313-22.
43. Pellegrini M, Cheng JC, Voutila J, Judelson D, Taylor J, Nelson SF, Sakamoto KM. (2008) Expression profile of CREB knockdown in myeloid leukemia cells. BMC Cancer 8: 264.
44. Zhou J, Bi C, Chng W, Cheong L, Liu S, Mahara S, Tay K, Zeng Q, Li J, Guo K, Tan CPB, Yu H, Albert DH, Chen C. (2011) PRL-3, a metastasis associated tyrosine phosphatase, is involved in FLT3-ITD signaling and implicated in anti-AML therapy. PLoS One 6: e19798.
45. Biernaux C, Loos M, Sels A, Huez G, Stryckmans P. (1995) Detection of major bcr-abl gene expression at a very low level in blood cells of some healthy individuals. Blood 86: 3118-22.
46. Aten JA, Stap J, Krawczyk PM, van Oven CH, Hoebe RA, Essers J, Kanaar R. (2004) Dynamics of DNA double-strand breaks revealed by clustering of damaged chromosome domains. Science 303: 92-5.
47. Wu Z, Irizarry RA, Gentleman R, Martinez-Murillo F, Spencer F. (2004) A model-based background adjustment for oligonucleotide expression arrays. Journal of the American Statistical Association 99: 909-917.
48. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JYH, Zhang J. (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 5: R80.
49. Dai M, Wang P, Boyd AD, Kostov G, Athey B, Jones EG, Bunney WE, Myers RM, Speed TP, Akil H, Watson SJ, Meng F. (2005) Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. Nucleic Acids Res 33: e175.
50. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. (2002) The human genome browser at UCSC. Genome Res 12: 996-1006.

Appendix A4

Optimizing the design and specificity of CRISPR genome engineering tools

The microbial nuclease system called CRISPR (**C**lustered **R**egularly **I**nterspaced **S**hort **P**alindromic **R**epeats) has been recently discovered and characterized for application in eukaryotic genome engineering. In prokaryotic systems, CRISPR functions as an immune system that confers resistance to exogenous plasmids and phages. Briefly, the system functions in bacteria by incorporating short segments of foreign DNA ("spacers") into the bacterial genome at CRISPR loci (which encode *cas*, or CRISPR-associated nuclease or helicase proteins as well as non-coding RNA elements that confer specificity of nucleic acid cleavage). RNA expressed from the CRISPR locus is then processed into small RNAs that guide the *cas* proteins to silence specific exogenous genetic material (e.g., by cleaving homologous dsDNA sequences).

It has been recognized that this microbial system could be exploited for engineering targeted cleavage (and subsequent deletion, or recombination) in the genomes of human (or other eukaryotic) cells. Methods have now been developed to use CRISPRs as a powerful tool for genetic perturbation, alongside existing zinc-finger proteases and transcription activator-like effector nucleases (TALENs).[1] A single RNA chimera (a synthetic sgRNA "spacer") with complementarity to a desired target genomic sequence can be used to reprogram sequence specificity for the *Cas9* endonuclease (part of the type II CRISPR system). The target sequence for *Cas9* is determined by the sgRNA spacer, and further specificity is conferred by the requirement that the target sequence be followed in the genome by a proto-spacer-associated motif (PAM) sequence (which for *Cas9* is 'NRG').

Until very recently, the genome-wide specificity of Cas9 CRISPR systems had not been characterized; it was unknown to what extent cleavage could occur at off-target sites sharing only partial homology with the sgRNA spacer (e.g. sites with 1-5 mismatched base pairs). We developed

a bio-informatics tool to predict, score, and assay all genome-wide off-target sites in the human genome for a given sgRNA spacer – and also select spacers that are optimized for a low off-target burden. This tool was used in a recent publication[2] to characterize the degree of cleavage at predicted off-target sites, and is now also available as an online web tool (http://www.genome-engineering.org/crispr/; web interface built by Benjamin Holmes) for the experimental community to optimize the choice of CRISPR spacers across a target region.

What does the CRISPR Design tool do?

The **CRISPR Design computational pipeline** optimizes the choice of sgRNA within a user's target sequence. The goal is to minimize total off-target activity across the human genome. For each possible sgRNA choice, the tool identifies all off-target sequences (preceding either NAG or NGG PAMs) across the human genome that contain up to 5 mismatched base-pairs. The cleavage efficiency at each off-target sequence is predicted using an experimentally-derived weighting scheme (based on data described here[2]). Each possible sgRNA is then ranked according to its total predicted off-target cleavage; the top-ranked sgRNAs represent those that are likely to have the greatest on-target and the least off-target cleavage.

In addition, this tool facilitates automated reagent design for CRISPR construction, primer design for the on-target Surveyor assay, and primer design for high-throughput detection and quantification of off-target cleavage via next-generation sequencing.

What kind of information does the tool take in?

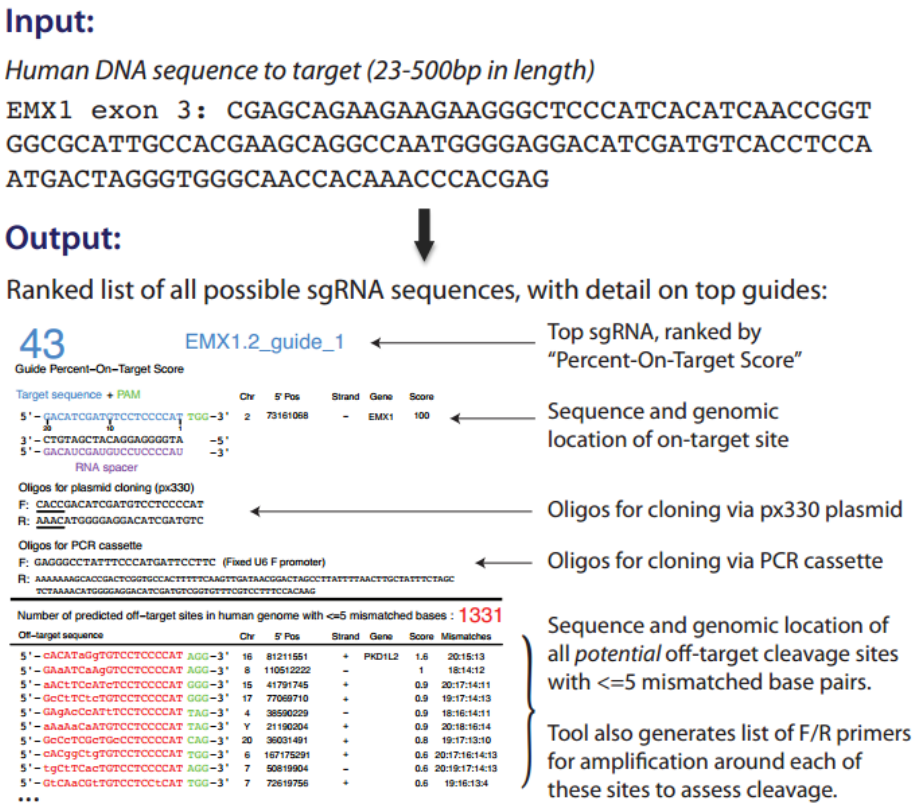
The CRISPR Design tool takes in three simple inputs from the user:

1. User's target region input sequence (between 23-1000bp, e.g. the genomic sequence of an exon)
2. A target identifier (e.g. the name of the gene user would like to target)
3. User's e-mail address

What kind of output does the tool generate?

1. A PDF file "_CRISPR_design.pdf" containing a summary of all possible sgRNAs, ranked in order of increasing predicted off-target cleavage.

2. An Excel spreadsheet "_CRISPR_design.csv" containing the sequence, location, and primers required for design of every possible sgRNA. This spreadsheet also contains information on all predicted off-target sequences for each sgRNA: the number of mismatched base-pairs, and the chr, pos, strand, and gene (if any) at which the off-target cleavage may be predicted to occur.
3. An Excel spreadsheet "_off_target_primers.csv" containing a set of F and R primer pairs that could be used to amplify ~90-130bp around the on-target and all off-target sequences predicted for every possible sgRNA. These primers are optimized for construction of a next-generation MiSeq sequencing library.



An example of the PDF output pages we developed is shown above; such a page is generated for all possible sgRNA sequences identified within the user's input sequence. Users can peruse the top 10 predicted off-target sites in this visual output, as well as all off-target sites in the spreadsheet output. Different sgRNAs have different properties: some have very few (e.g. 1-2) high-scoring off-targets and a tail of very weak off-targets, others have no high-scoring off-targets but a very large number of weak off-target sites. Some users might prefer the former scenario, in which clones could

be screened for a single off-target site; others may prefer the latter scenario in which the overall probability of off-target cleavage may be lower. Some users may be especially interested in minimizing protein-coding off-target sites. The tool is designed to enable users to design CRISPR reagents that are best-suited for their application.

We have used this tool to optimize the design of CRISPRs targeting both the human and rat genome (the latter because members of our laboratory wanted to perform engineering in the rat pancreatic insulinoma INS1E cell line).

References

1. Barrangou, R. RNA-mediated programmable DNA cleavage. *Nature Biotechnology* **30**, 836-8 (2012).
2. Hsu, P.D. *et al.* DNA targeting specificity of RNA-guided Cas9 nucleases. *Nature Biotechnology* 1-8 (2013).doi:10.1038/nbt.2647

Appendix A5

List of all manuscripts to which contributions were made

Below is a list of all manuscripts (published and in preparation) to which contributions were made during the course of completing this thesis.

2013

Agarwala V*, Flannick J*, Sunyaev S, and Altshuler D. "To what extent can empirical data place bounds on the genetic architecture of complex human diseases?" *Accepted in Nature Genetics* (2013).

Moutsianas L*, **Agarwala V***, Fuchsberger C, Flannick J, Rivas M, Gaulton K, The GoT2D Consortium, McVean G, Boehnke M, Altshuler D, and McCarthy M. "The power of gene-based rare variant methods to detect disease-associated variation and test hypotheses about complex disease." *Under review* (2013).

Agarwala V, Hartl C, Rasmussen M, Morris A, Wright J, Flannick J, Mahajan A, Alston J, Fontanillas P, Ward L, Kang HM, The GoT2D Consortium, Kellis M, Boehnke M, McCarthy M, and Altshuler D. "Systematic characterization of allelic architecture at the non-coding type 2 diabetes locus 9p21 using haplotype analysis in complete sequence data, imputation, and functional testing." *Manuscript in preparation* (2013).

Wang S, **Agarwala V**, Flannick J, Altshuler D, GoT2D Consortium, Hirschhorn J. "Simulation of Finnish population history, guided by empirical genetic data, to assess power of rare variant association tests." *Manuscript in preparation* (2013).

Flannick J, Beer N, Bick A, **Agarwala V**, Molnes J, Gupta N, Burtt N, Florez JC, Meigs JB, Taylor H, Lyssenko V, Irgens H, Fox E, Burslem F, Johansson S, Trimmer J, Newton-Cheh C, Tuomi T, Molven A, Wilson JG, O'Donnell CJ, Kathiresan S, Hirschhorn J, Njølstad P, Rolph T, Seidman J, Gabriel S, Cox D, Seidman C, Groop L, and Altshuler D. "Assessing the phenotypic effects in the general population of rare variants in genes for a dominant Mendelian form of diabetes." *Under review* (2013).

Hsu PD, Scott DA, Weinstin JA, Ran FA, Konermann S, **Agarwala V**, Li Y, Fine EJ, Wu Xuebing, Shalem O, Cradick TJ, Marraffini LA, Bao G, and Zhang F. ‘‘DNA targeting specificity of RNA-guided Cas9 nucleases." *Nature Biotechnology* (2013).

Ran FA*, Hsu PD*, Wright J, **Agarwala V**, Scott DA, and Zhang F. "A CRISPR Toolbox for Genome Engineering." *Accepted in Nature Protocols* (2013).

2012

Engreitz JM*, **Agarwala V***, Mirny LA. "Three-Dimensional Genome Architecture Influences Partner Selection for Chromosomal Translocations in Human Disease." *PLoS ONE* 7(9): e44196 (2012).

2011

Shea J, **Agarwala V**, Philippakis A et al. "Comparing strategies to fine-map the association of common SNPs at chromosome 9p21 with type 2 diabetes and myocardial infarction." *Nature Genetics* 43, 801-5 (2011).

Jordan, D.M., Kiezun, A., Baxter, S.M., **Agarwala, V.,** Green, R.C., Murray, M.F., Pugh, T., Lebo, M.S., Rehm, H.L., Funke, B.H., and Sunyaev SR. "Development and Validation of a Computational Method for Assessment of Missense Variants in Hypertrophic Cardiomyopathy." *American Journal of Human Genetics,* Volume 88, Issue 2, 183-192 (2011).