THE UNIVERSITY OF CHICAGO

BEYOND THE LOW-HANGING FRUIT: LEVERAGING SUMMARY DATA IN HUMAN POPULATION AND STATISTICAL GENOMICS

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF HUMAN GENETICS

BY
MICHAEL CHARLES TURCHIN

CHICAGO, ILLINOIS
DECEMBER 2017

Table of Contents

List of Figures

List of Tables

ACKNOWLEDGMENTS

I am truly grateful for the time I have had here at the University of Chicago. I was fortunate enough to be accepted to the Department of Human Genetics PhD program, and to be present during many PIs' tenures as part of the department. I have been able to grow as a scientist, a collaborator, a mentor, a student, and as a colleague during my time here. And these developments are due in no small part to the multitude of amazing and exceptional people I have had the pleasure to work with these past six years.

I would like to first thank my committee of John Novembre (Chair), Anna Di Rienzo, and Xin He. John has provided countless advice throughout my years here, and has always been available for a quick chat. He was always a go-to source for me in regards to the inner-workings of academia, for which I will always be grateful. Anna has continually provided a high standard for the quality of work both for the smaller details as well as the larger picture of a project. Her breadth, and depth, of knowledge has also been a boon both in, and out of, our committee meetings. Xin has always provided a valuable and novel perspective throughout my committee meetings. He was an amazing complement to my committee and always encouraging throughout the process.

I would also like to thank Jonathan Pritchard, whose lab I was a member of during his time at the University of Chicago. I was fortunate enough to be part of multiple projects in Jonathan's lab, one of which made it to publication. Jonathan's curiosity and determination were incredibly rewarding to be a part of, and helped set a pro-

ductive tone to the beginning of my graduate school career. Jonathan also cultivated a diverse and talented group of researchers, many of whom helped impact my career. I am indebted to having been welcomed to Jonathan's lab.

I would also like to thank the following administrators in Human Genetics, who helped smooth the many edges of graduate school during my time here: Justin Osadjan (HG Program Manager), Erin Brady (HG Program Manager), Candice Lewis (HG Program Manager), Iwona Niekras, Holly McGuinness, Vikki Webster, and Anita Williams-Logan. I would also like to thank Sue Levinson (Genetics, Genomics, and Systems Biology) for always taking care of HG students as well, Lucia Rothman-Denes (Molecular Genetics and Cell Biology) for managing the Genetics and Regulation Training Grant, and Diane Hall (BSD Executive Administrator) for always helping resolve issues.

I would like to thank the following individuals who over the years were part of the Cummings Life Sciences Building 4^{th} Floor. These include individuals from the Stephens, Novembre, He, Pritchard, Przeworski, Di Rienzo, and Ober Labs: Hussein Al-Asadi, John Blischak, Peter Carbonetto, Kushal K Dey, David Gerard, Joyce Hsiao, Mengyin Lu, Kevin Luo, Jean Morrison, Lei Sun, Sarah Urbut, Gao Wang, Wei Wang, Siming Zhao, Xiang Zhu, Ester Pantaleo, Raman Shah, Zhengrong Xing, John Zekos, Roger Pique-Regi, Allegra Petti, Bryce van de Geijn, Carolyn Jumper, Anil Raj, Yair Field, Ziyue Gao, Ellen Leffler, Daniel Matute, Wynn Meyer, Aarti Venkat, Keerthi, Sylvia Kariuki, Mark Rappel, Ben Peters, Evan Koch, Joel Smith, Joe Marcus, Dan Rice, Arjun Biddanda, Nick Knoblauch, and Shengtong Han.

In particular I would also like to acknowledge the following CLSC 4^{th} floor members, all of whom helped guide me as an early-stage graduate student: Jacob Degner*, Timothée Flutre, Audrey Fu, Bryan Howie, Ida Moltke, Heejung Shim, Xiang Zhou, Shyam Gopalakrishnan*, Graham McVicker, Amir Kermany, and Joe Maranville (* – special acknowledgement for always making time for my many questions).

I would also like to acknowledgement my graduate school cohort (HG & GGSB): Nicholas Banovich, Choongwon Jeong, Katie Igartua, Carolyn Jumper, Zach Weiler, Jason Pitt, and Bryan Lenneman. We were very fortunate to have the group that we did during our years together. I was glad our first year was solidified by the never-ending Facebook message thread. And I was glad to share many memories along the way years after. You were all people who I could confide in and invariably made graduate school that much more possible to finish. Thank you.

I will also thank my family, whom has always been unquestionably supportive throughout the decade-plus of work leading to this moment. I am fortunate to have parents who never once questioned my decision to pursue academic research. In fact, my parents have always encouraged the journey that I have gone on. Graduate school comes with many challenges, and I am glad that constantly justifying my choices over the phone once a week was never one of them. Thank you guys for always being there throughout the process.

And lastly, I want to acknowledge and thank my PI, Matthew Stephens. Matthew has given me an invaluable experience and opportunity by being part of his lab; I have grown immeasurably and leave with a stronger sense of my scientific-self. As a

researcher who exists along the border of multiple fields, having supportive mentors I think is particularly important – it is easier to fall back into a pre-ordained mold, rather than continually push the envelope of what type of researcher you can be. Matthew has always been encouraging and supportive as I spent my graduate school career attempting to understand where I fit in. And I believe my success in this endeavor is vastly due to his guidance, patience, and many conversations. I will always be grateful for what Matthew has taught me, both as a scientist and as an individual.

ABSTRACT

A main goal of human genetics is to connect naturally occurring genotypic variation to naturally occurring phenotypic variation. Over the past three decades, human geneticists accomplished this through collecting cohorts of individuals and genotyping various kinds of DNA markers. Historically, human geneticists have been limited by the number of samples they could analyze, the breadth of genomic regions they could target at any one time, and the phenotypes available to them. However, these restrictions are now being rapidly mitigated – advances in genotyping and sequencing technologies have made a vast majority of the genome accessible to researchers, and the increasing affordability of these technologies has made creating larger cohorts much more feasible. Additionally, a wide range of gene-regulatory phenotypes is now available to human geneticists, diversifying the questions researchers can pursue. As a result, the amount of data being produced now is unprecedented. But with this new wealth of information comes new challenges. The first waves of results from new technologies have been mixed, producing exciting, novel findings but also disappointment – for many human traits there remains a large unexplained proportion of trait heritability. While some researchers have responded to this disappointment by collecting more samples and applying newer technologies, others have focused on how to make better use of the data we already have. In this dissertation, I present three projects that follow this latter theme, each attempting to use summary information from pre-existing data or first-stage analyses to better answer a research question. In Chapter 2 I present results from a study using pre-existing data (Exome Sequencing Project and 1000Genomes) to identify whether the deleterious mutational load in dif-

ferent between human populations. We show that despite the differing demographic histories of European-Americans and African-Americans these populations have similar mutational loads. We also show through simulations and theoretical predictions that this result is expected. In Chapter 3 I present the first stage results from a two-phase HIV candidate gene-exome sequencing study. In this study we sequenced the exomes of ∼1,300 genes each with *a priori* experimental evidence of being functional related to HIV. Using ∼750 individuals from the Multicenter Aids Cohort Study (MACS), we show an enrichment for marginal association signals among between candidate genes and HIV-Acquisition. We also show a proposed target list being used to conduct follow-up genotyping in the full MACS cohort; this list was heavily informed by summarizing the results from our first-stage analysis. Lastly, in Chapter 4 I present results from applying a Bayesian multivariate genome-wide association study method (bmass) on 13 different publicly available GWAS datasets. bmass runs on univariate GWAS summary statistics and identifies not only new significant variants but also the underlying multivariate patterns driving these results. We show that for many of the datasets analyzed we find additional variants, and we also display the multivariate patterns discovered for each dataset as well.

CHAPTER 1

INTRODUCTION

1.1 Human Genetics – The Genetic Basis of Complex Traits and Association Studies

One of the major goals of human genetics is to identify the underlying genetic basis of physical traits[20, 39, 114, 181, 126, 124, 182, 96, 155, 64, 4, 8, 165]. Often in molecular genetics, this is accomplished by experimentally altering genetic material and identifying downstream physical changes (e.g. broadscale mutagenesis[132, 186, 206, 197], breeding schemes[37, 38, 221, 185], and site-directed mutagenesis[122, 205, 137, 106, 92]). By using these kinds of approaches, researchers can demonstrably show a direct link between a genetic region and the physical trait that region affects. However, human geneticists cannot utilize these methods – experiments like this cannot be conducted on human populations. Instead of directly perturbing genetic material, human geneticists take alternative approaches to connect genotype to phenotype. Often these alternative approaches involve taking advantage of the vast, natural genetic diversity observed in global human populations. Instead of experimentally altering DNA, human geneticists attempt to identify patterns of genetic variation that align with patterns of phenotypic variation. And one fundamental branch of methods that uses this strategy is known as association studies.

An association study compares specific genetic regions from across individuals in an

attempt to find strong correlations between genotypes and a phenotype of interest. For example, consider the following scenario: we are interested in discovering genomic regions involved with susceptibility to HIV. To accomplish this, we collect a large number of individuals that either have HIV or do not. We then extract genetic information from specific regions, such as a handful of genes, from every individual we collected. With this information in hand, we then attempt to identify both a) genetic mutations fluctuating in our population of individuals and b) mutations that strongly differentiate our two subset of individuals. Imagine we sequenced gene A in all our individuals, and at basepair position 10 we see a mutation from an adenine (A) to cytosine (C). However, we additionally observe that all our infected individuals have the C allele and all our uninfected individuals have the A allele; this observation would strongly suggest that having either the C or A allele at this position is somehow associated with HIV-susceptibility. And furthermore, such a result would implicate gene A as being somehow involved with HIV-susceptibility. This is the type of finding an association study aims to discover, with the next steps likely being to identify **how** this mutation in gene A alters HIV-susceptibility (e.g. via pathway analysis, overlaying with epigenetic information, using human cell or mouse models, etc...).

Association studies have been conducted by human geneticists for over two decades now, with the scale of both the number of individuals used and number of loci being targeted increasing over time. Early on in the 1990s, human geneticists – limited by costs and available technology – would focus on just a handful of loci in cohorts of modest sizes (<10,000 individuals). In this approach, known as 'candidate gene

studies'[78, 161, 149, 123], researchers would select and sequence loci that appeared to be phenotypically relevant based on previous experimental work. For more Mendelian traits (phenotypes with few causal loci, each with large effect sizes), candidate gene studies had some success, such as with genetic modifiers of Huntington's disease[83] and some cancer susceptibility genes[46, 166]. However for more complex traits (traits with many causal loci, each with modest effect sizes) such as diabetes and bipolar disorder[109, 63, 66], candidate gene studies were largely inconclusive. Often, initial results from candidate gene studies in complex traits would fail to replicate in follow-up cohorts[101, 26, 140]. And as a result, researchers began to question whether the candidate gene approach was well suited for complex traits – at the sample sizes being used, it was possible that the lack of replication should in fact be expected[182]. So as human geneticists reconciled how to adapt association studies for complex traits, a shift in research efforts began to occur, with a new focus on increasing both sample sizes and the number of loci being tested per study.

1.2 The Build Up of Genome-Wide Association Studies

To adapt association studies to complex traits, human geneticists needed to both target more regions of the genome per study as well as increase the number of samples being used per study. In part to accomplish these goals (as well as due to what was technologically possible at the time), researchers began focusing on a specific type of genetic mutation – single nucleotide polymorphisms, or SNPs. SNPs are single basepair mutations, often (though not restricted to) diallelic, that are actively

segregating in a given population. SNPs are present across the human genome, appearing on average at least once per 1kb[194]. Additionally, because the human genome is structured in haplotype blocks[35, 70, 232], SNPs can effectively 'tag' most regions of the genome by representing these blocks[107, 27, 90]. Therefore human geneticists hypothesized that SNPs might be well suited as genetic markers that could represent most of the entire genome.

To transform SNPs into genome-wide, association-study suitable markers, researchers made two important SNP-related advancements. First, human geneticists discovered and mapped a large number of SNPs in multiple human populations. The most prominent of these efforts was the HapMap project[96] – a large-scale study in the mid 2000's that originally focused on individuals of European, African, or Chinese descent. Through this work, the first release produced a genome-wide map of 1,007,329 high-quality SNPs[97] (two follow-up releases added to this total [99, 98]). Second, biotechnology companies (such as Illumina and Affymetrix) then used these SNP maps to produce relatively affordable genotyping platforms known as 'SNP-chips'[167, 13, 133, 204]. Optimized to target SNPs that best tagged the variously sized haplotype blocks in the human genome, SNP-chips were able to genotype anywhere from 250,000 to >1,000,000 SNPs at a time. Therefore these SNP-chips made generating genome-wide information across a large number of individuals feasible for the first time, thus opening the door for the association studies originally envisioned. With both these SNP maps and technologies in hand, human geneticists began to conduct what would become known as genome-wide association studies, or 'GWAS'.

There was perceptible excitement in the human genetics community with the dawn of applying GWAS to complex traits. Because most of the genome was tagged by these SNP-chips, it was anticipated that the majority of complex trait genetic determinants would be discovered. Early returns from GWAS included finding hundreds of SNP associations for not only various disease states such as Type 2 Diabetes, Crohn's Disease, and Bipolar Disorder[237, 44, 14, 61, 100], but for physical characteristics such as height[128], BMI[203], lipids levels[216], glucose levels[48], and more[89, 87, 225] as well. Additionally, some early GWAS discoveries overlapped known drug-targets as well, such as SNP associations between LDL and *HMGCR* (a gene target for cholesterol-reducing statins)[216, 239] and between rheumatoid arthritis and *PADI4 & IL-6R* (targets of the inhibitor BB-Cl-amidine and suppressor Tocilizumab, respectively)[213, 54, 111]. By 2012 alone over 2,000 individual SNP-trait associations cumulatively had been discovered[226]. Clearly, using GWAS did begin to reveal the previously elusive genetic architecture of complex traits.

However, as the number of SNP associations increased, human geneticists began to realize it was not just a matter of identifying these associations, but also interpreting them. Multiple issues began to arise with interpreting SNP associations, but most prominent among them was an issue involving estimated trait heritabilities. For many complex traits studied, researchers already had established values of these traits' heritability – the proportion of phenotypic variance explainable by genetic effects. For example height is known to have a heritability of roughly 80%[128]. However, when researchers began to estimate trait heritabilities using all the SNP associations that had been discovered thus far, they found their estimates to often

be far below what they expected. For example, using the GWAS-significant SNPs from the initial round of height GWAS, heritability was only estimated at ∼12%; other traits that returned low heritability estimates using initial sets of GWAS-significant SNPs included 20-25% in LDL and HDL levels and 20-25% in Crohn's Disease as well (see Table 1 of Lander 2011; an estimate of ∼60% in Type 1 Diabetes using GWAS-significant SNPs was one of the few closer values). This discrepancy between well-established heritabilities and newly estimated SNP-based heritabilities became a growing concern and earned the moniker of the "Missing Heritability" problem[144, 200, 145, 249].

1.3 The Follow-Up to Genome-Wide Association Studies

The "Missing Heritability" publications acted as important checkpoints for human geneticists. Researchers began to question if they truly were discovering the most important aspects of complex trait genetic architecture like they had anticipated. GWAS had indeed made many important genetic discoveries, but it seemed like researchers still had much more work to do to properly understand complex trait genetic architecture. Human geneticists began to debate both what they were possibly missing from their current models of complex trait architecture as well as which GWAS follow-up directions should first be explored. Here we discuss two of these follow-up avenues in particular as they pertain to projects presented in this dissertation.

A fundamental aspect of GWAS and the SNP-chips commonly employed was that they targeted 'common genetic variation' – SNPs whose minor allele frequencies ('MAF') were often >1%. This choice was based on the theoretical perspective that complex traits would be most influenced by common genetic variation, known as the 'Common Disease Common Variant (CDCV)' hypothesis[180, 190]. This perspective suggested most complex traits are determined by the cumulative modest effects of weakly penetrant variants; and since these mutation effect sizes are modest, they should be mostly segregating in human populations as close-to-neutral alleles (i.e. insignificantly impacted by the effects of negative selection). This is in contrast to the 'Common Disease Rare Variant (CDRV)' hypothesis[175, 190], which posits that most complex traits are determined from the large effects of strongly penetrant variants; and under this scenario, such large-effect variants would often be found as 'rare genetic variation' (MAF <1%; such as being new mutations from recent human demographic processes, or continually suppressed by mutation-selection balance). Therefore human geneticists began returning to this debate, suggesting that the missing heritability could possibly be found in this rare variation space unexplored by GWAS.

Conducting rare-variant studies however requires different technologies than GWAS, technologies that were unfeasible in the mid-2000's but had begun to mature by the mid-2010's. To explore genome-wide rare variation, human geneticists need either whole-exome or whole-genome sequencing[193, 125, 79], methods that return full exonic or genomic DNA sequence information. By retrieving the actual sequence information researchers can directly discover any variant present in their group of

samples. Doing so reveals the entire allele frequency spectrum, which would have been less feasible for SNP-chips to accomplish (e.g. rare variation is more population and individual specific[81]). Therefore with sequencing technologies now becoming more affordable and accessible, human geneticists began to conduct sequencing-based association studies targeting rare variation. Early returns suggest that at least for some complex traits, rare variation may indeed play a substantial role, such as with autoimmune diseases[183, 153, 15, 25] and neurodevelopmental disorders[156, 75, 178, 41]. Yet there is already debate on how best to design sequencing studies and construct rare-variant tests[79, 129, 236, 170], thus indicating the large amount of work this still needs to be done.

Another major avenue of post-GWAS follow-up has been to simply increase the power of association studies. By increasing power, GWAS should hypothetically continue to find more significant common SNP associations. This is important in the context of the "Missing Heritability" problem since previous work has shown, at least for height (a particularly polygenic trait), that continuing to include more significant SNPs in calculations of heritability leads to larger estimates[243, 241]. For example going from height GWAS results from 2010 to height GWAS results from 2014 increased estimates of heritability from $\sim$12% to $\sim$20% (by going from using 180 GWAS SNPs to 697 GWAS SNPs)[128, 241]. Additionally, it was shown in the 2014 study that by incorporating all studied SNPs agnostic of their association p-values, heritability was estimated at $\sim$60%[241] (an earlier study also using height GWAS SNPs found similar results[243]). Therefore, at least in the case of particularly polygenic traits, the "Missing Heritability" may simply be found by using a much larger number of

SNPs than human geneticists originally anticipated. And so to find these many additional SNPs, there is a need to increase GWAS power.

There are multiple ways to increase power in GWAS. One common approach is simply to use more samples. The larger number of height SNPs previously mentioned between the 2010 and 2014 results was primarily due to discovery set sample sizes increasing from ∼130,000 to 250,000[241]. With the advent of both genome sequencing consortiums and BioBanks such as ExAC[131] and the UKBioBank[211], we are already seeing community-wide efforts to increase available GWAS-ready samples. Another common approach to increase power in GWAS is to use more advanced statistical methods – such as tests that use multivariate approaches. Often human geneticists conduct GWAS using univariate approaches, i.e. methods that test genotypes against only a single phenotype at a time. However, an alternative setup to this is to use multivariate approaches, methods that tests genotypes against multiple phenotypes simultaneously[105, 196, 244]. It has been shown that under a number of biological scenarios multivariate approaches actually increase power in GWAS[105, 247, 72] compared to univariate approaches, such as when multiple traits are linked to a genetic variant and even when only a single trait is linked to a genetic variant. These latter scenarios occur when multiple phenotypes are strongly correlated; even if only one trait is actually linked to a variant of interest, if there are additional correlated traits, including them as covariates will regress out shared, non-genetic effects (see Figure 1a in Stephens 2013).

In fact there are multiple aspects of GWAS that lends the framework to multivariate

approaches. First, it is not uncommon for researchers to measure multiple traits in their cohorts of interest[48, 95, 230, 240, 241, 138, 198, 11]. Often this is a product of both multiple traits having related biological importance as well as the experimental logistics of measuring said traits (e.g. it may be particularly easy, or even necessary at times, to measure multiple phenotypes). For example GWAS on blood lipid levels commonly analyze low-density lipoproteins, high-density lipoproteins, total triglycerides, and total cholesterol; a major reason why these four traits in particular are analyzed is because they are all simultaneously measured by commonly-used blood lipid panels[240]. Furthermore, we expect all four of these traits to be correlated since they are highly related to one another biologically.

Second, GWAS results can more effectively be interpreted with the addition of complementary genomic information, and multivariate approaches provide a useful context. Whether by determining multiple traits are jointly associated (or explicitly determining additional traits are in fact not associated), multivariate approaches can help narrow down which are the most relevant biological pathways. For example, imagine you have a SNP you determined to be associated with BMI via a univariate analysis. If by doing additional multivariate analyses you determine the SNP is jointly associated with type 2 diabetes as well, this provides additional context that affects your interpretation – possibly now you pay more attention to insulin-resistance or other pancreatic pathways than you would have previously. Indeed, recent work has shown both extensive sharing of genetic effects (Pickrell et al. 2016[171] analyzed 42 different traits and found 341 loci that affect pairs of phenotypes) and unassociated, correlated phenotypes powering association signals

(Stephens 2013[207] analyzed the Global Lipids Consortium[216] data and found multiple, best SNP-association models including unassociated, correlated phenotypes) across complex traits, suggesting there are many opportunities to glean additional context for interpreting GWAS results.

Here we present three projects that all attempt to do work relevant in the post-GWAS era. Two of the projects presented here specifically deal with alternative GWAS study designs (rare-variant studies) and alternative GWAS methods (multivariate tests). Additionally, all three projects in this dissertation attempt to incorporate summary information from next-generation methods, an important concept as the human genetics community continues to generate new data. Such as what was found with GWAS, new technologies in the human genetics community often produce both exciting results and new problems. And while follow-up directions to address these new problems commonly include both analyzing the data more thoroughly (e.g. developing better models and methods) and generating more data (e.g. using newer technology or increasing sample sizes), there is a tendency to take the latter path of research. Often newer technologies begin getting increased attention well before the problems and concerns from the previous generation of technologies are well addressed. Therefore, we attempt in each of these projects to go deeper with pre-existing data, either by reusing previously generated data for a different question, informing the next stages of a project using generated data, or reanalyzing already published data.

In the first project (Chapter 2) we use results from the exome sequencing project

('ESP') to address an ongoing question in human genetics, whether the deleterious mutational load is different between European-Americans and African-Americans. This question has been tackled multiple times before, but with at times ambiguous or conflicting results between studies[139, 28, 112, 158, 215]. One possibility for the contradictory results is different types of genetic data being used to answer this same question. Here, we employ summary allele frequency information from the ESP as well as simulations to try and reconcile previous work and produce a more consistent answer.

In the second project (Chapter 3) we attempt to identify genes that are significantly associated with HIV-Infection using a candidate target approach. GWAS in the field of host HIV genetics has produced significant associations but only in a few regions of the genome[57, 24, 58, 169, 151]; in an attempt to narrow the focus of genomic interrogation, we sequence the exomes of genes that had prior experimental evidence of being related to HIV-host interactions. Additionally because of the lack of success from evaluating common variation through GWAS, we emphasize analyzing the rare variation we identify through our sequencing results. We take a two-phase study design approach where we first use a subset of an HIV cohort to evaluate our candidate gene list, and then second we summarize and use these results to prioritize targets for follow-up genotyping in the full cohort.

And in the third project (Chapter 4) we extend a previously published framework for Bayesian multivariate association studies (via a software package 'bmass') and apply it to multiple publicly available datasets. As previously mentioned, GWAS

are well-suited for multivariate analyses; however, multivariate approaches are still infrequently used. While there are likely multiple reasons for this, we focus on one issue in particular – interpretation of multivariate results. Often with multivariate methods, it is difficult to determine how much any single phenotype contributes to a signal of association. For example, early multivariate approaches such as MANOVA or SNPTEST[147] would evaluate the model 'all phenotypes are associated'. While this model would increase power for identifying associations, a positive result did not automatically indicate all phenotypes were equally driving the association signal. Imagine we compute a significant p-value for the model SNP **g** is associated with Height and BMI; from this result alone we do not know how much either Height or BMI contribute to the association signal. If we look at the univariate models of **g** associated with either Height or BMI and find **g** is significantly associated with Height but not BMI, how does this affect our interpretation? Should these results temper our findings that **g** is associated with Height and BMI? These questions only become more complex as the number of phenotypes and models grow exponentially. We aim to address this issue in two important ways: we provide a user-friendly framework that explicitly tests all possible models and also indicates the relative support each model has. With this framework one can then take a quantitative look at the above example to see how much stronger the signal of association becomes from **g** associated with Height to **g** associated with Height and BMI. And we show the utility of this framework by running it on multiple publicly available datasets and providing the results.

Overall, we aim to show multiple projects that move beyond typical GWAS study

designs and GWAS datasets. We aim to show how various summary metrics from these datasets can be used to produce more results and additional biological insight. And we aim to provide evidence that taking the time to think more deeply about the data we already have is a worthwhile endeavor while we also enthusiastically employ new technologies.

CHAPTER 2

THE DELETERIOUS MUTATION LOAD IS INSENSITIVE TO RECENT POPULATION HISTORY

Yuval B. Simons[1,*], Michael C. Turchin[2,*], Jonathan K. Pritchard[2,3,4†] and Guy Sella[1,5,6,†]

[1]Department of Ecology, Evolution, and Behavior, The Hebrew University of Jerusalem

[2]Department of Human Genetics, The University of Chicago

[3]Howard Hughes Medical Institute

[4]Departments of Biology and Genetics, Stanford University

[5]Department of Ecology and Evolution, The University of Chicago

[6]current address: Department of Biological Sciences, Columbia University

[*]These authors contributed equally.

[†]To whom correspondence should be addressed: pritch@stanford.edu, gsella@math.huji.ac.il.

2.1 Abstract

Human populations have undergone dramatic changes in population size in the past 100,000 years, including recent rapid growth. How these demographic events have affected the burden of deleterious mutations in individuals and the frequencies of disease mutations in populations remains unclear. We use population genetic models to show that recent human demography has likely had little impact on the average burden of deleterious mutations. This prediction is supported by two exome sequence datasets showing that individuals of west African and European ancestry carry very similar burdens of damaging mutations. We further show that for many diseases, rare alleles are unlikely to contribute a large fraction of the heritable variation, and therefore the impact of recent growth is likely to be modest. However, for those diseases that have a direct impact on fitness, strongly deleterious rare mutations likely do play an important role, and recent growth will have increased their impact.

2.2 Introduction

Recent work has highlighted the impact of demographic history on the distribution of human genetic variation. Deep sequencing studies have identified huge numbers of very rare variants in human populations, the consequence of explosive population growth in the past five thousand years[34, 148, 68, 112, 158, 215]. Additionally, Europeans and east Asians have a greater fraction of high-frequency variants compared to Africans, likely due to an ancient bottleneck of non-African populations

[233, 229, 113, 84, 215].

Given these observations, it is natural to ask whether recent demographic history has impacted the burden of genetic disease in modern human populations[139, 28, 68, 112]. Keinan and Clark[112] recently hypothesized that "Some degree of genetic risk for complex disease may be due to this recent rapid increase in the number of rare variants in the human population". A second important question concerns the relative importance of rare and common variants in causing disease[176, 55, 74]. If much of the genetic variation underlying disease is due to rare variants, then this could help to explain the so-called "missing heritability" of complex traits, and imply that mapping approaches based on deep sequencing will be essential for the dissection of complex traits[146].

2.3 Results

To address these questions, we analyzed a theoretical model with a large number of bi-allelic sites, each subject to two-way mutation, and natural selection against one of the alleles (see Methods for details). We studied three types of demographic models thought to be relevant for human populations: (i) a bottleneck; (ii) exponential growth starting from a constant-sized population; and (iii) a complex demographic model for African Americans (including rapid recent growth) and European Americans (including two bottlenecks followed by growth) inferred by Tennessen *et al.*[215]. The main features of the Tennessen model are similar to other recent

models[189, 229, 84] while using a larger data set for parameter estimation. Our main results focus on selection against semi-dominant (i.e., additive) alleles in which the three genotypes have fitnesses 1, $1 - s/2$ and $1 - s$, respectively; and selection against recessive alleles with genotype fitnesses 1, 1, and $1 - s$. The effects of demography in these two models are qualitatively representative of those over the range of dominance coefficients (Supplement Note, Section 2.4). In addition to simulation results shown here, further results and detailed theoretical analysis for all our key results are provided in the Supplement.

2.3.1 The impact of demographic changes on individual load

We focus first on the impact of demographic changes on individual load – that is, we want to understand whether demographic history has impacted the burden of deleterious variation carried by a typical individual in a population. Individual load is directly related to the number of deleterious alleles carried by an individual, or for recessive mutations to the number of homozygous sites per individual (see the Methods and Supplement for further details).

Figure 1 illustrates the impact of a bottleneck and population growth on the numbers of deleterious variants with strong selection (s=1%). As expected, these demographic events have a major impact on the number and frequency spectra of deleterious variants: the bottleneck causes a decrease in the total number of segregating sites in a population due largely to loss of rare variants, while the mean frequency of alleles

that survive increases. Meanwhile, exponential growth causes a rapid increase in the number of segregating sites due to a major influx of rare variants, but a consequent drop in the mean frequency at segregating sites. But despite these dramatic shifts in the overall frequency spectrum, the impact on genetic load – namely, the mean number of deleterious variants per individual and thus the average fitness – is much more subtle.

In the semi-dominant case, the load is essentially unaffected by these demographic events (Figures 1C and 1D). With growth, the increased number of segregating sites is exactly balanced by a decrease in mean frequency (and conversely for the bottleneck), so that the number of variants per individual stays constant. This kind of balance is predicted by classic mutation-selection balance models18, and can be shown to hold for general changes in population size, provided that selection is strong and deleterious alleles are at least partially dominant (Supplementary Note, Section 2.3).

The behavior of the recessive model is more complicated (Figures 1E and 1F). In the bottleneck model, the mean number of deleterious variants per individual drops by 60% as a result of the bottleneck. This is due to the loss of rare alleles. However, during the bottleneck, some deleterious alleles drift to higher frequencies[219, 139], contributing disproportionately to the number of homozygotes. This causes a transient increase in the number of deleterious homozygous sites per individual – i.e., the recessive load. Meanwhile, population growth has a less pronounced effect on recessive variation, leaving the mean number of deleterious alleles per individual

unchanged, but causing a slight decrease in load.

More generally, the manner in which demography affects load varies with the degree of dominance and the strength of selection (Figure 2, Supplementary Note, Section 2 & Supplementary Table 1). The behavior of these models can be classified into three selection regimes (strong, weak and effectively neutral). In the strong selection case, i.e., where selection is much stronger than drift (approximately $s \geq 10^{-3}$ for semi-dominant mutations), deleterious variants are extremely unlikely to fix, and virtually all of the genetic load is due to segregating variation. In this range, we infer that human demography has had no impact on semi-dominant load (and more generally for mutations with at least some dominance component), and small effects on recessive load.

The weak selection case–where drift and selection have comparable effects–is more complex, as fixed alleles may contribute appreciably to load, and steady state load depends on population size[141]. However, the approach to steady state is very slow, being limited both by the time to fixation (on the order of $4N$ generations) and by the mutational input (on the order of $\frac{1}{2Nu}$ generations). For both the semi-dominant and recessive cases, population growth is too recent to have substantially decreased the load. Recent growth increases the input of new deleterious mutations, but this effect is counterbalanced by the fact that the new deleterious mutations are proportionally rarer. The bottleneck in Europeans is estimated to have occurred farther in the past and at much lower population sizes[215] (Supplementary Figure 1), allowing it to have more effect. In this case, the increase in drift causes segregating deleterious

alleles to increase in frequency, sometimes reaching fixation, and results in a slight increase in load (Supplementary Figure 2). The out-of-Africa bottleneck should thus lead to a slight increase of load in Europeans, most notably for recessive sites.

Finally, in the effectively neutral range – where selection has negligible effects on the population dynamics – segregating variation contributes negligibly and hence the load does not change with demography. Thus, across all three selection regimes, recent human demographic history is likely to have had virtually no impact on genetic load at partially dominant sites, and only weak effects at recessive sites.

2.3.2 Analysis of exome data

To test these predictions, we analyzed two recent data sets of exome sequences from individuals of west African and European descent. Previous work comparing load in different populations has produced conflicting conclusions depending on the dataset, choice of measures and functional annotations. For example, Lohmueller *et al.*[139] reported that there is "proportionally more deleterious variation in European than in African populations". Similarly, Tennessen *et al.*[215] found that European Americans had more non-reference genotypes when they used a conservative classification of deleterious sites, but observed the opposite when using a more liberal classification of sites (both observations were highly significant).

We first analyzed single nucleotide variant (SNV) frequency data from a recent exome sequencing study of 2,217 African Americans (AAs) and 4,298 European Americans

(EAs) sequenced at 15,336 protein coding genes by Fu*et al.*[68] (allele frequencies available from the NHLBI GO Exome Variant Server). Additionally we analyzed exome data from 88 Yoruba (YRI) and 81 European (CEU) individuals collected by the 1000 Genomes Project[217].

To test whether there are differences in load between individuals of west African and European descent, we considered the average number of derived alleles per individual at putatively deleterious segregating sites. For this purpose, a site is considered to be segregating if and only if it is variable within the combined sample of both populations. This definition ensures that the derived counts are comparable across populations. Under a semi-dominant model, the number of derived alleles increases monotonically with the segregating genetic load. Thus, any difference in average load between populations would be apparent as a difference in the mean number of derived alleles per individual. Here, we focused on an equivalent measure that also facilitates comparisons across different types of sites: namely, the mean derived allele frequency within functional classes. Note that the mean derived allele frequency is simply equal to the number of derived alleles per individual divided by twice the number of segregating sites in that class, and so any difference in the mean number of derived alleles per individual will also be a difference in mean derived frequencies. For sites that are either neutral or semi-dominant, our model predicts that the mean derived allele frequency should be virtually identical in Africans and Europeans (Supplementary Note, Section 3 & Supplementary Figure 3). At recessive sites, we expect a slight increase in mean derived frequency in Africans compared to Europeans (Supplementary Figure 3), but overall we expect any differences to be

small.

Functional predictions of SNVs were obtained from PolyPhen2, a method that uses sequence conservation and structural information to infer which non-synonymous changes are most likely to have functional consequences[6]; see Supplement Table 2 for similar analyses with other functional prediction methods. When using the functional predictions we observed a strong bias: SNVs where the genome reference carries the derived allele are much more likely to be classified as benign than SNVs where the reference allele is ancestral – this is true even when we control for the overall population frequency (Supplementary Figure 4). Hence our analysis incorporates a correction to account for this bias; we also obtained very similar results using a separate set of unpublished human-independent PolyPhen scores kindly provided by the Sunyaev lab (Supplementary Table 4).

Figure 3 summarizes the results for the data of Fu *et al.* As expected, the mean allele frequency declines with increasing functional severity[215], from 2.8% at non-coding SNVs to 0.6% at probably-damaging SNVs, implying that there is selection against most SNVs with predicted damaging effects. More striking, however, is that within each of the five functional categories, the mean allele frequencies – and hence the numbers of derived alleles per individual – are essentially identical in the two populations, despite the very large size of the data sets ($p>.05$ for all five comparisons). Results for the 1000 Genomes Project data are qualitatively similar: we find no significant differences between YRI and CEU in the numbers of derived alleles per individual in any functional category (Supplementary Table 5).

In summary, these observations are consistent with our model predictions that load should be very similar in these populations. Our conclusions likely differ from previous studies partly because earlier studies used measures that are related to load but are also sensitive to other differences between the populations being compared (e.g., the number of neutral segregating sites and the frequency spectrum) and partly due to the reference bias in functional annotations accounted for here (see Supplementary Note, Section 3). We note that David Reich, Shamil Sunyaev and colleagues have recently made similar observations regarding load in different populations (personal communication).

2.3.3 The impact of demography on the genetic architecture of disease susceptibility

Although population size changes have had little impact on the average load carried by individuals, growth has greatly increased the number of rare variants in populations. So do rare variants play a greater (and substantial) role in the genetics of disease as a result of recent growth (Figure 4)? Given the differences in population history, do higher frequency variants play a greater role in Europeans and Asians than in Africans? The answers to these questions are of practical importance because different study designs may be needed to identify rare variants[176, 146, 74, 218].

To study this, we computed the contributions of different allele frequencies to the heritable phenotypic variation among individuals in the population, namely $x(1 -$

$x)f(x)/2$, where $f(x)$ is the probability that a derived allele is at frequency x given the demographic model and selection coefficient. These distributions show the fraction of genetic variance for a disease that is contributed by alleles below frequency x, for the simplest case where the loci underlying a trait all have the same effect size, the same selection coefficient, and are semi-dominant (see Supplementary Note, Section 4). In practice, we anticipate that variants underlying a given disease would have a variety of selection coefficients and effect sizes, in which case the overall distribution would be an appropriately weighted mixture of distributions for different selection coefficients. Note that in this model, we consider the proportional contribution of variants at different frequencies and thus, these results should hold regardless of the number of loci underlying variation in the trait.

Analysis of this model reveals several interesting points. For effectively neutral, or for weakly deleterious sites (Figure 4A), only a small fraction of the total variance comes from very rare alleles: although there are many rare alleles, each one contributes very little to population variance and individual load. The same is true for recessive variation across almost the entire range of selection coefficients (Supplementary Note, Section 4.2 & Supplementary Figure 5). Likewise, if we assume that the frequency density $f(x)$ follows the frequency spectrum observed at all non-synonymous sites classified as "probably damaging"[6] then, under the same model, it is still only a modest fraction of the genetic variance that is due to rare alleles (Figure 4B; c.f. ref. [215]). Meanwhile, in all of these cases an Out-of-Africa bottleneck would increase the contribution of intermediate frequency alleles to the genetic variance (Figure 4A-C): e.g., at probably damaging sites 62% of the variance in EAs is contributed by

alleles with minor allele frequency above 10% compared to only 49% in AAs.

It is only for the case of strong, dominant selection that very rare variants ($< 0.1\%$) become important (Figure 4C and 4D). For example, for a selection coefficient of 1%, most of the variation is rare and arose within the recent exponential growth phase. As a result, the contribution of extremely rare variants is much greater than it would have been in the absence of growth: e.g., in AAs and EAs, 80%, and 65% of the variance is due to alleles below frequency 0.1%, compared to just 25% in the constant population model.

Of course in practice, the genetic variants that contribute to a complex trait likely have a range of selection coefficients (s) and a range of effect sizes (a) on the phenotype in question (Supplementary Note, Section 4.3). When there is a mixture of selective coefficients, what can we say about the relative importance of rare and common variants? To answer this, the critical issue is to model the relationship between a and s [55, 108]. To illustrate this, we consider two extreme cases: (1) a is independent of s, namely, the trait itself has little effect on fitness but specific variants could have fitness consequences due to pleiotropic effects on other phenotypes; and (2) a is proportional to s – likely most relevant for traits with a direct impact on fitness such as early-onset diseases or diseases affecting fertility. Figure 4E shows the expected genetic variance per site as a function of s under these two models. When a is independent of s model, we would expect weakly selected mutations to contribute most of the variance because they have the same average effect on the trait but can drift to higher frequencies. But the reverse occurs in the model where a increases

with s: highly deleterious, rare mutations will have a greater contribution to variance because their increased effect size outweighs their lower frequencies.

Many traits presumably lie between these two extreme cases. To study how demography affects genetic architecture across this range, we consider a second model. We assume that the heritable variance in a trait is due to a mixture of weakly ($s = 0.0002$) and strongly ($s = 0.01$) selected mutations and we vary the correlation between selection on a variant and its effect on the trait (see Methods for details). Figure 4F shows how the contribution of rare alleles to genetic variance changes with the correlation between the selection coefficient and effect size. As can be seen in the case with constant population size, the contribution of rare variants becomes substantial only when the variants' effects on fitness and on the trait are highly correlated (presumably because the trait itself is strongly coupled with fitness). While growth affects the frequencies of strongly selected alleles regardless of the correlation, it will have a substantial effect on the genetic architecture of a trait only for traits in which strongly selected alleles contribute substantially to variance. In this case, we see that the recent growth greatly amplifies the contribution of rare alleles to the variance. A similar argument implies that the Out-of-Africa bottleneck should substantially increase the contribution of intermediate frequency alleles to the variance, unless the effects of variants on fitness and on the trait are highly correlated, in which case rare alleles will still dominate.

2.4 Conclusion

While recent demographic events have had well-documented effects on the frequency spectrum of SNVs in modern populations, we find that these events have had negligible impact on the average burden of mutations carried by individuals. Moreover, we conclude that although there are large absolute numbers of rare variants, they do not necessarily contribute a large fraction of the genetic variance underlying complex traits. An earlier paper from one of the present authors (Pritchard, 2001[176]) also discussed the possible role of allelic heterogeneity and rare variants in disease using a model that is closer to the independent s model here. While the earlier model is not exactly comparable to our present work, the overall results are broadly consistent, as the bulk of the genetic variance was predicted to be due to variants that would not be considered rare by modern standards. To summarize, it is only for diseases that are primarily due to strongly deleterious mutations that we can expect much of the variance to be due to rare alleles: these will likely tend to be diseases that are tightly coupled to fitness.

2.5 Acknowledgements

This work was supported by grants from the National Institutes of Health (MH084703, GM083228), the Israel Science Foundation (grant # 1492/10), and the Howard Hughes Medical Institute. MT was supported in part by NIH grant T32 GM007197. Thanks to Molly Przeworski and Graham Coop for comments and discussions; to David Reich and Shamil Sunyaev for helpful discussions and generous input regarding the interpretation of PolyPhen 2; to Ivan Adzhubey for human-independent PolyPhen scores; to Josh Akey for assistance in accessing data; and to Josh Akey, Adam Siepel, and an anonymous reviewer for comments on the manuscript.

2.6 Competing Interests.

The authors declare that they have no competing financial interests.

2.7 Online Methods

This section provides a summary of our methods; a complete version may be found in the Supplementary Information.

2.7.1 Model

Our basic model starts by considering selection at a single site. We use the standard bi-allelic diploid model with two-way mutation, viability selection, drift and, in some cases, migration[29]. Specifically, we assume there are two possible alleles at each site: normal (N) and deleterious (D). An N allele mutates to the D allele with probability u per gamete, per generation and the reverse mutation occurs with probability v. Unless noted otherwise, we assume that mutation is symmetric, i.e., u=v. The absolute fitness of the three genotypes NN, ND and DD are 1, 1 - hs and 1 - s, respectively, where s ¿ 0 and h $\geq$ 0. We focus on semi-dominant (h=1/2) and fully recessive (h = 0) selection because these two cases exhibit the full range of qualitative behaviors, with selection acting primarily on heterozygotes when h¿1/2 and only on homozygotes when h=0. Allele frequencies in the next generation follow from Wright-Fisher sampling with these viabilities, sometimes with migration, and the population size and migration rates vary according to the demographic scenario considered.

We assume that fitness is multiplicative across sites, and that there is linkage equi-

librium among sites. Under these assumptions, the evolutionary dynamics at each site are independent from all other sites. In practice, linked selection is likely to have negligible effects on differences between populations because, to a first approximation this reduces the effective population size at a given site by similar proportions regardless of demographic history and these effects are thought to be modest in humans (e.g., ref. [152]).

2.7.2 Demographic scenarios

We consider three demographic scenarios. The most detailed is the Out-of-Africa demographic model for African-Americans (AA) and European-Americans (EA) estimated by Tennessen et al.[215] (Supplementary Figure 1A). The model includes the Out-of-Africa split of European ancestors, changes in population size before and after the split (specifically, a severe bottleneck in Europeans following the split and recent rapid growth in both Europeans and Africans) and migration between the populations after the split. Finally, the model includes recent admixture between the populations, which we include in our simulations only when we compare our results to data from AAs.

We also study two simpler demographic scenarios (Supplementary Figure 1B&C). To understand the effects of recent explosive growth of human populations, we use a simple model of exponential growth from a population of constant size and similarly, to investigate the effects of the bottleneck in Europeans at the Out-of-Africa split,

we consider a simple model of a bottleneck where population size instantaneously changes to a lower value at which it stays constant until it instantaneously reverts back to its original size.

2.7.3 Simulations

For each demographic scenario, we run simulations of a single site for the semi-dominant and recessive cases and vary the selection coefficient such that the strength of selection ranges from effectively neutral to strong. Each run begins with one of the two alleles fixed, where the proportion of runs that start with each allele is given by the expectation at equilibrium. A burn-in period of $\geq 10N$ generations with constant population size N follows in order to ensure an equilibrium distribution of segregating sites. The initial state is defined as ancestral and the other state as derived; the derived and deleterious allele frequencies are recorded at the end of the simulation. The code is written in C++ and is available upon request. (See Supplementary Note, Section 1 & Supplementary Figures 6-8.)

2.7.4 Load

Genetic load is defined as the relative reduction in average fitness caused by deleterious alleles, compared to the maximum absolute fitness [29]. In our model, the maximal absolute fitness is equal to 1, allowing us to directly consider differences in average fitness in populations with different demographic histories. Given our model,

the average fitness function can be written as

$$\bar{W} \approx \exp(-\sum_{j=1}^{M} l(h_j, s_j))$$

where

$$l(h, s) \equiv 2hsE(pq) + sE(q^2) = s(2hE(q) + (1 - 2h)E(q^2)), \tag{2.1}$$

relates the quantities at a locus with load, p and q are the beneficial and deleterious allele frequencies at a locus ($p+q = 1$) and h_j and s_j are the dominance and selection coefficient at locus j. For a model with a single site and $s \ll 1$, $l(h, s)$ coincides with the definition of load. For more than one site, load is a simple function of the sum over $l(h, s)$'s. For brevity, we therefore refer to $l(h, s)$ as load.

2.7.5 Change in load

To assess whether there has been a change in load due to demography, we consider the difference between load at the present time and the load before recent demographic events. Specifically, in the exponential and bottleneck models the reference time is before the change in population size and in the Tennessen model the reference time is the split between the African and European populations. (See Supplementary Note, Section 2, Supplementary Figures 2, 9-20 & Supplementary Table 1.)

2.7.6 Data Analysis

We used data from Fu et al. (2012)[68] and from the 1000 Genomes Project[217]. Allele frequency estimates from Fu *et al.* are available from the NHLBI GO Exome Variant Server (`http://evs.gs.washington.edu/EVS/`). These provide estimates of the derived allele frequencies at exonic SNVs in European- and African-Americans (EA and AA). Variants with allele frequencies 0 or 1 in both EA and AAs were excluded. 1000 Genomes Project vcf files (Phase 1 Version 3) were downloaded from the official 1000 Genomes public server. YRI and CEU individuals with (at least) exome sequencing coverage were extracted from the original .vcf files (88 YRI individuals and 81 CEU individuals). 7 YRI individuals, chosen at random, were removed to match sample sizes between YRI and CEU. Variants that were fixed for either allele in both populations were removed. Any variant that was not an SNV or did not contain ancestral allele information was also dropped.

The ANNOVAR suite of scripts [235] was used to obtain functional predictions for each SNP from each of four prediction methods: PolyPhen2 [6], SIFT [121], LRT [32] and MutationTaster [191]. We observed a strong reference bias in the functional classifications for all four prediction methods: sites at which the reference genome carries the derived allele are much more likely to be classified as benign than are sites where the reference is ancestral; this is a very strong effect even when we control for the true population frequency in a very large sample (Supplementary Figure 4), and hence does not simply reflect the tendency for common alleles to be less functional. We therefore treated the functional designations at sites where the genome reference

is derived as unreliable. To deal with this problem we used a simple procedure to estimate the probability that each reference-derived site would have been classified as damaging had the reference allele been ancestral (conditional on the overall population frequency). Specifically, we binned SNVs by overall population frequency in the full sample and, for each bin, we determined the fraction of reference-ancestral sites in each functional category. For SNVs in that bin that are reference-derived, we treated those fractions as estimates of the probability that these SNVs would have been in each functional category had they instead been reference-ancestral. Next, to estimate the mean derived allele frequency (DAF) for each functional category, we summed across all sites in that category that were reference ancestral, and added a contribution from all sites that were reference-derived, weighted according to the estimated probability that the site would have been in the relevant functional category if it had been reference-ancestral. We also provide supplementary results in which we used a new unpublished version of PolyPhen's PSIC scores that are calculated in a human-independent (i.e., unbiased) manner and obtain qualitatively similar results. We thank Ivan Adzhubey and Shamil Sunyaev for pre-publication access to these.

We calculated mean derived frequencies within functional categories, and the corresponding standard errors (calculated as SD(DAF)/Sqrt(#sites). Individual-level counts for the 1000 Genomes data simply counted the numbers of derived alleles per individual within a functional class (note that there are no missing genotypes in this data set as these have been imputed). For each population and functional category we estimated the standard deviation of the mean number of derived alleles

per individual by bootstrapping across sites. This is more appropriate than computing the standard error directly from the distribution of derived allele counts across individuals, as the latter method ignores variation in the evolutionary process. Note that because we are working with mean allele counts or frequencies, these analyses are unaffected by linkage disequilibrium or Hardy Weinberg disequilibrium (which may affect variances but not means).

Note that our analysis effectively uses the derived allele count as a proxy for the deleterious allele count. Hence, there will be a low rate of misclassification at weakly selected sites for which the deleterious allele is ancestral. However this does not change the qualitative predictions about patterns of differences between populations and we expect the number of derived alleles to have a monotonic relationship with the number of deleterious alleles. Specifically, for sites that are either neutral or semi-dominant, we predict that this measure should yield virtually identical counts in AAs and EAs (Supplement Note, Section 2 & Supplementary Figure 20). At recessive sites, our model predicts slight differences (Supplementary Note, Section 2), but overall we expect these differences to be negligibly small. Note that when SNVs are defined within populations as in some previous papers, these simple predictions do not hold.

2.7.7 Models for variance

We consider how the relationship between the effects of mutations on fitness and a trait affect genetic architecture. For that purpose, we calculate the expected contribution of mutations to the heritable variation in a trait. We assume an additive trait and that the fitness effects of mutations are semi-dominant. At a site with selection coefficient s, the expected contribution to the variance from deleterious alleles below frequency ω is therefore

$$V_\omega(s) = \frac{1}{2} C E(a^2|s) \int_0^\omega f(x|s) x(1-x) dx, \tag{2.2}$$

where $E(a^2|s)$ is the expectation of the squared effect size, $f(x|s)$ is the probability of the deleterious allele being at frequency x (without conditioning of the site being segregating, i.e., including $x = 0$ and 1) and the C is a proportion coefficient (cf. Supplementary Note, Section 4.1). A site's expected contribution to variance is $V_1(s)$ and the proportional contribution from variants below frequency ω is $\Theta_\omega(s) \equiv \frac{V_\omega(s)}{V_1(s)}$; Note that while $V_1(s)$ depends on the relationship between selection coefficients and effect sizes, $\Theta_\omega(s)$ does not. When all sites are considered jointly, denoting the input of mutations with selection coefficient s by $\mu(s)$, the expected proportion of variance from deleterious alleles below frequency ω is

$$\Theta_\omega = \frac{\int_s \mu(s) V_1(s) \Theta_\omega(s) ds}{\int_s \mu(s) V_1(s) ds}. \tag{2.3}$$

As an illustration, we consider a simple model in which we vary the correlation

between selection on variants and their effects on a trait. We assume that half of the newly arising mutations have a weak selection coefficient $s_w = 0.0002$ and half have a strong selection coefficient of $s_s = 0.01$. For strongly selected mutations, the effect size on the trait, a, is chosen to be cs_s with probability $\frac{1}{2}(1+p)$ and cs_w with probability $\frac{1}{2}(1-p)$, where c is a positive constant and $0 \leq p \leq 1$; correspondingly, for weakly selected mutations the effect size is chosen to be cs_w with probability $\frac{1}{2}(1+p)$ and cs_s with probability $\frac{1}{2}(1-p)$. In this model, the marginal distributions of selection coefficients and effect sizes do not depend on p, while the correlation between them is equal to p. To obtain Figure 4F we therefore varied p between 0 and 1. In Figure 4E, we consider the two extremes ($p = 0$ and 1).

2.7.8 URLs

The NHLBI GO Exome Variant Server, http://evs.gs.washington.edu/EVS; The 1000 Genomes public server, ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/.

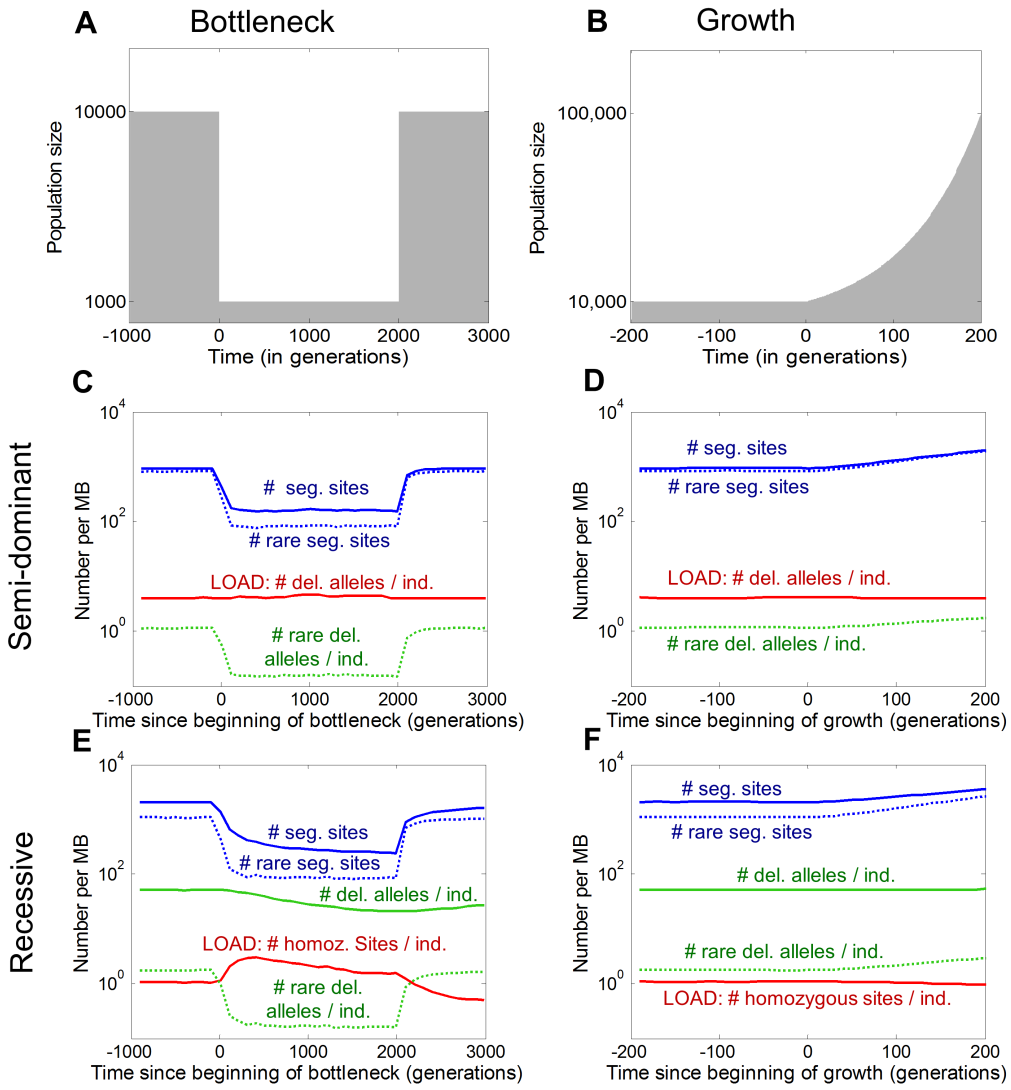

Figure 2.1: **Time course of load and other key aspects of variation through the course of a bottleneck (panels A, C, E) and exponential growth (panels B, D, F).** Each data line shows the expected number of variants, or alleles per MB, assuming semi-dominant mutations (panels **C**, **D**) or recessive mutations (panels **E**, **F**) with s=1% and mutation rate per site per generation=10^{-8}.
Versions of these plots with linear scales can be found in Supplementary Figures 13, 14, and 16.

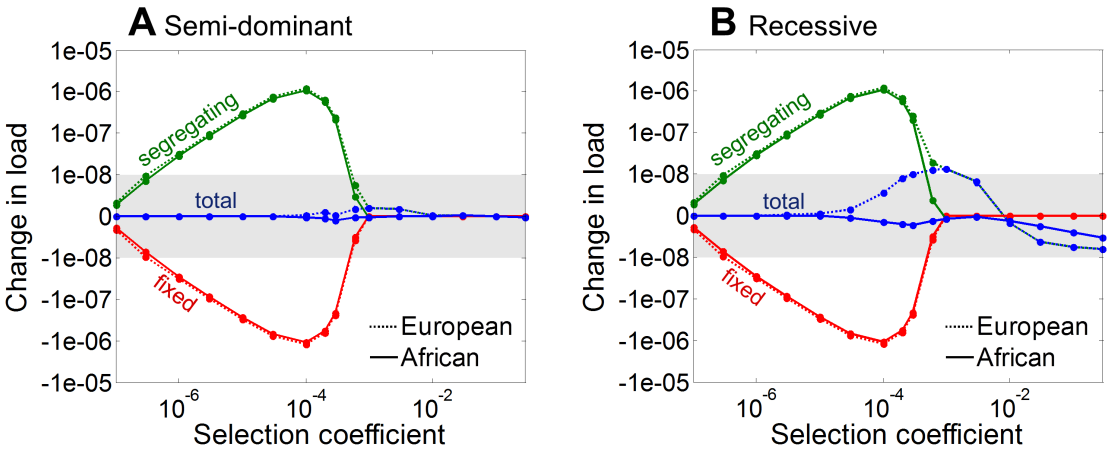

Figure 2.2: **Changes in load due to changes in population size during the histories of European and African Americans for (A) semi-dominant and (B) recessive sites.** The blue lines show the difference in total expected load per base pair of DNA sequence in the present day population compared to the ancestral (constant) population size, as a function of selection coefficient. The green and red lines show the difference in the amount of load due to segregating and fixed variants, respectively. As can be seen, there is more load due to segregating variation in modern populations, but this approximately cancels with reduced load to fixed sites as shown by the total load lines (blue). The y-axis scale is linear within the grey region and logarithmic outside.

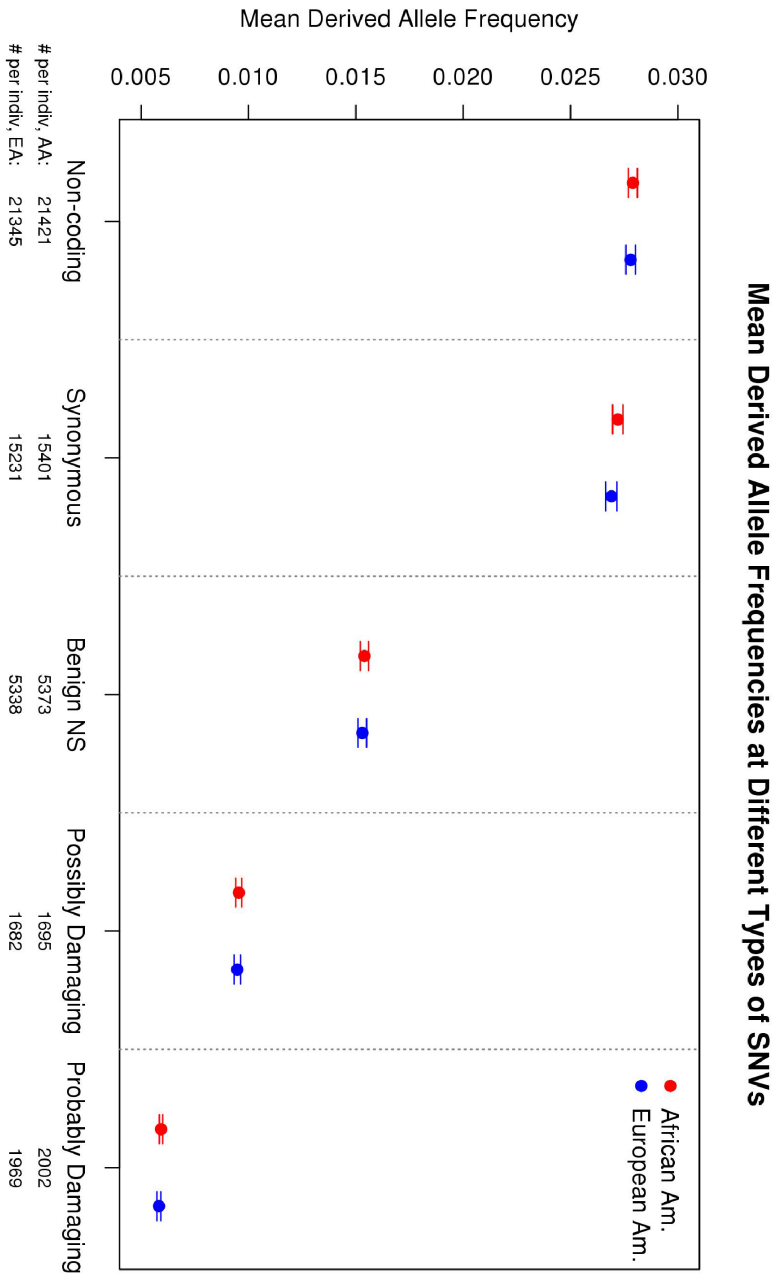

Figure 2.3: **Observed mean allele frequencies in African and European Americans at various classes of SNVs.**

Figure 2.3 (Cont.): **Observed mean allele frequencies in African and European Americans at various classes of SNVs.** The plot shows mean frequencies in each population, plus and minus two standard errors, using exome sequence data from Fu et al.[68]. Here a site is considered an SNV if it is segregating in the combined AA-EA sample of 6515 individuals. The functional classifications of sites are from PolyPhen2[6] with bias-correcting modifications. The AA and EA mean frequencies are essentially identical within all five functional categories ($p>0.05$).

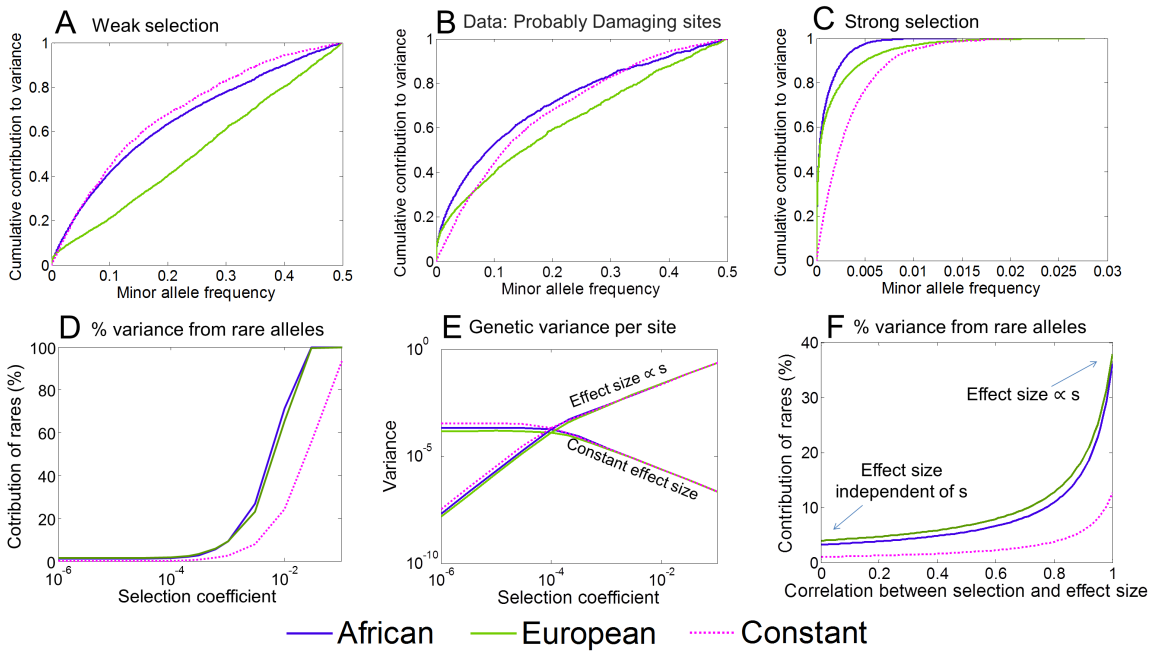

Figure 2.4: **Predicted effect of demography on the genetic architecture of disease risk.** All the plots assume an additive trait and, with the exception of (B), are based on simulations with semi-dominant selection under the Tennessen *et al.*[215] demographic model. Results for the constant population size model are also provided for comparison. The upper plots show the cumulative fractions of genetic variance due to alleles at frequency ¡ x, based on: (A) simulated data with weak selection (s =.0002); (B) assuming the observed frequency spectrum at 'probably damaging' sites[68, 6], where a constant population size of 14,474 and selection coefficient of 0.02% are used for comparison; and (C) simulated data with strong selection (s = .01). Panel (D) depicts the fraction of variance due to rare alleles (i.e., ¡ 0.1%) as a function of the selection coefficient; (E) shows the per-site contribution to variance as a function of the selection coefficient under two extreme models, with effect sizes that are either independent of s (constant) or proportional to s; (F) shows the expected fraction of the variance due to rare variants (i.e., ¡ 0.1%) as a function of the correlation between the selection on, and effect size of variants. Further details on the model are provided in the Methods.

2.8 Supplementary Methods

2.8.1 Model and Simulations

Our basic model considers selection at a single site. We use the standard bi-allelic diploid model with (in this order) two-way mutation, viability selection, drift and, in some cases, migration [29]. Specifically, we assume there are two alleles at a site: normal (N) and deleterious (D). An N allele mutates to the D allele with probability u per gamete, per generation and the reverse mutation occurs with probability v. Unless noted otherwise, we assume that mutation is symmetric, i.e., $u = v$. The absolute fitnesses of the three genotypes NN, ND and DD are 1, $1 - hs$ and $1 - s$, respectively, where $s > 0$ and $h \geq 0$. We focus on semi-dominant ($h = \frac{1}{2}$) and fully recessive ($h = 0$) selection because these two cases exhibit the full range of qualitative behaviors (with selection acting primarily on heterozygotes in one and only on homozygotes in the other), but we also consider the robustness of our findings to other dominance coefficients (section 2.8.2). Allele frequencies in the next generation follow from Wright-Fisher sampling with these viabilities, sometimes with migration, and the population size and migration rates vary according to the demographic scenario considered.

For each demographic scenario, we ran simulations of a single site for the semi-dominant and recessive cases and varied the selection coefficient such that selection ranges from effectively neutral to strong. For a given set of parameters, the number of runs was determined by requiring a sampling error of less than 2% in estimates

of the main summaries (e.g., the mean deleterious allele frequency and squared frequency). Error bars denoting estimates of one standard deviation around the mean are provided in all the graphs based on simulations, unless they are too small to be visible. Each run begins with one of the two alleles fixed, where the proportion of runs that start with each allele is given by the expectation at equilibrium. A burn-in period of $\geq 10N$ generations with constant population size N follows in order to ensure an equilibrium distribution of segregating sites. The initial state is defined as ancestral and the other state as derived; the derived and deleterious allele frequencies are recorded at the end of the simulation. The code is written in C++ and is available upon request.

Demographic scenarios. We consider three demographic scenarios. The most detailed is the Out-of-Africa demographic model for African-Americans (AA) and European-Americans (EA) estimated by Tennessen et al. [215] (Figure 2.5A). The model includes the Out-of-Africa split of European ancestors, changes in population size before and after the split (specifically a severe bottleneck in Europeans following the split and recent rapid growth in both Europeans and Africans) and migration between the populations after the split (see Figure 2.5A for details). Finally, the model includes recent admixture between the populations, which we include in our simulations only when we compare our results to data from AAs.

While the Tennessen et al. model was parameterized in a diffusion framework, i.e., in continuous time, Wright-Fisher simulations require discrete numbers of generations and individuals. We therefore divide the times by 25 years per generation (the

generation time that Tennessen et al. assume) and round the number of individuals associated with any of the parameters (e.g., growth) to the nearest integer. We implement migration by sampling alleles from the local population with probability $1 - m$ and from the other population with probability m each generation.

We also study two simpler demographic scenarios. To understand the effects of recent explosive growth of human populations, we use a simple model of exponential growth with parameters matching those of the African population in the Tennessen et al. model (see Figure 2.5B for details). For the purpose of analysis, this scenario is sometimes extended by adding a period with constant population size after growth ends. Similarly, to investigate the effects of the bottleneck in Europeans at the Out-of-Africa split, we consider a simple model of a bottleneck with parameters matching those of the European bottleneck in the Tennessen et al. model (see Figure 2.5C for details). Here, we sometimes extend the period after the reduction in population size to study longer-term equilibration to reduced population sizes.

Validating the simulation. We used two approaches to check the validity of the simulations. For a constant population size, we compared the frequency spectra from simulations with those expected under the diffusion approximation (cf. [53]) for the neutral case as well as for several semi-dominant and recessive selection coefficients (Figure 2.6). We note that obtaining similar frequency spectra implies that simpler summaries, such as the number of segregating sites under neutrality or the average deleterious allele frequency at mutation-selection balance, will also be similar.

For the more elaborate Out-of-Africa demographic model, we compared the minor

allele frequency spectrum from neutral simulations with the spectrum observed at non-coding sites in Fu et al. [68] We cosider non-coding sites for this purpose as these are assumed to be under the least selection (Figure 2.7). In their Figure 2A, Tennessen et al. find a close agreement between the observed spectra and a diffusion approximation under their demographic model. We find close agreement of our neutral simulations to data from both AAs and EAs and the slight differences that we do find are similar to those in their Figure 2A [215].

Sensitivity to mutation rate. Unless noted otherwise, we follow Tennessen et al. [215] in using a mutation rate of $u = 2.36 \cdot 10^{-8}$ per bp per generation. Given that recent estimates suggest a lower mutation rate (e.g. Kong et al. [116], Sun et al. [212]), we examine here the sensitivity of our simulation results to this assumption. We find the derived allele frequency spectrum to be extremely robust, remaining essentially unchanged when we double or halve the mutation rate (Figure 2.8A). As expected, the number of segregating sites and the number of sites fixed for the derived allele increase (linearly) with the mutation rate (Figure 2.8B). The increase in the number of sites fixed for the derived allele follows from the increased rate of fixation in the burn in period (akin to fixations that occur between the ancestor of humans and chimpanzees and the Out-of-Africa split). Thus, assuming a different mutation rate will affect some of our quantitative results. Notably, if the mutation rate in humans is indeed lower than the one we use, as recent estimates suggest, the proportion of segregating sites would be lower, resulting in an even smaller effect of recent demographic history on load than our analysis suggests (see section 2). Our qualitative finding of a negligible effect on load is unchanged. Moreover, our results

concerning the effects of recent demography on genetic architecture derive from the frequency spectrum and therefore are unaffected.

2.8.2 The effects of demography on load

We assume that fitness is multiplicative across sites and that selected sites are at Linkage Equilibrium (LE). The absolute fitness of individual i can then be written as

$$W_i = \prod_{j=1}^{M} w_{i,j},$$

where the product is taken over the M sites contributing to fitness and $w_{i,j}$ is the contribution of site j, which depends on the genotype of the individual and on the selection and dominance coefficients at that site. Given LE, the contributions of sites to the expected fitness in the population are independent and therefore

$$E(W_i) = \prod_{j=1}^{M} E(w_{i,j}) \approx \exp(-\sum_{j=1}^{M}(2h_j s_j p_j q_j + s_j q_j^2)),$$

where p_j and q_j are the frequencies of the normal and deleterious alleles at site j. We note that the approximation applies for strong selection because the frequency q_j is small, as well as for weak selection because then the selection coefficient is small. Finally, taking an expectation over evolutionary realizations (which is equivalent to an expectation over many sites with the same parameters in a single realization)

yields

$$E(W) \approx \exp(-\sum_{j=1}^{M}(2h_j s_j E(p_j q_j) + s_j E(q_j^2))). \tag{2.4}$$

The latter expression relates the population dynamics at a site with the overall reduction in fitness.

Genetic load is defined as the relative reduction in average fitness caused by deleterious alleles, calculated as

$$L = \frac{W_{max} - \bar{W}}{W_{max}},$$

where W_{max} is the fitness of an individual without deleterious alleles and $\bar{W}$ is the average fitness [29]. Denoting the terms associated with a single site in Equation 2.4 by

$$l(h, s) \equiv 2hsE(pq) + sE(q^2) = s(2hE(q) + (1 - 2h)E(q^2)), \tag{2.5}$$

the fitness function can be rewritten as

$$E(W) \approx \exp(-\sum_{j=1}^{M} l(h_j, s_j)).$$

This form emphasizes that the reduction in fitness caused by a single site generally depends on the first two moments of the deleterious allele frequency. Specifically, in the semi-dominant model, it depends only on the first moment

$$l(\frac{1}{2}, s) = sE(q),$$

and in the recessive model it depends only on the second

$$l(0, s) = sE(q^2).$$

Moreover, this form shows that $l(h, s)$ provides a natural additive measure for the expected reduction in fitness caused by a site.

Throughout the manuscript we therefore use $l(h, s)$ as our measure for the contribution of a site to load. For a model with a single site, it coincides with the definition of load, as $E(L) = l(h, s)$. For more than one site,

$$E(L) \approx 1 - \exp(-\sum_{j=1}^{M} l(h_j, s_j)).$$

Given that in our model, the load from all sites is a simple function of the sum of $l(h, s)$ across sites, for brevity, we refer to $l(h, s)$ as load.

With a constant population size, the load exhibits three standard dynamic regimes depending on the scaled selection coefficient (Figure 2.9): (i) An effectively neutral regime, in which $\alpha = 2Ns \ll 1$ and the effects of selection are negligible compared to drift; (ii) a weak selection (or nearly neutral) regime, in which $\alpha = 2Ns \approx 1$ and the effects of selection and drift are comparable; (iii) a strong selection regime, in which $\alpha = 2Ns \gg 1$ and selection dominates over drift.

In what follows our analysis is divided according to these three regimes. When the population size changes, the boundaries between regimes are affected. Moreover,

the rate at which the equilibrium for a new population size is attained depends on the summary of the data considered. We consider summaries for segregating sites, e.g., the proportion of segregating sites and the allele frequency at these sites, and summaries for fixed sites, e.g., the proportion of sites fixed for the deleterious allele (which we call fixed state). Specifically, we are interested in the effects of demography on the contribution of segregating and fixed sites to load, which we refer to as fixed and segregating load, and in their sum, which we refer to as total load. We consider the behavior of these statistics for the two simple demographic models, which together allow us to understand all qualitative behaviors exhibited under the more detailed Tennessen et al. model (2.10). For these demographic models, we primarily consider two modes of inheritance (semi-dominant and recessive).

To simplify our theoretical analysis, we make several reasonable assumptions about the parameters of the model. For brevity, we focus on the case with symmetric mutation ($u = v$) and, because we are considering human populations, we assume that the population mutation rate per site is small, i.e., that $\beta = 2Nu \ll 1$. We also assume that the selection coefficient is small, i.e., $s \ll 1$. A summary of our analyses are presented in Figure 2.10 and Table 2.1. A detailed description of the behavior in each regime follows.

The effectively neutral regime

When selection is negligible compared to drift, the behavior of deleterious alleles is well approximated by that of neutral alleles. As the properties of neutral alleles (e.g., the proportion of segregating sites and frequency spectrum) in models with constant and varying population sizes have been studied exhaustively (e.g., [214, 93, 231]), here we focus only on the implications concerning load.

First, we consider how load depends on the selection coefficient at equilibrium for a constant population size. If deleterious alleles behave like neutral ones, the first two moments of the deleterious allele frequency distribution do not depend on the selection coefficient and therefore the load is proportional to the selection coefficient (see Eq. 2.5). This explains the linear relationship between selection coefficient and load shown in Figure 2.9.

At equilibrium, load depends negligibly on the population size. Using the diffusion approximation for the stationary deleterious allele frequency distribution [53], the expansion of the load to first order in α and β yields

$$l(h,s) = \frac{s}{2}(1 - \frac{1}{2}\alpha - 2(1-2h)\beta).$$

Thus, as long as $\beta \ll 1$ and $\alpha \ll 1$, the load is well approximated by $s/2$ regardless of the population size and dominance coefficient (hence the similarity in load for the semi-dominant and recessive cases in Figure 2.9). Intuitively, this follows from the fact that the great majority of sites are fixed, and because selection is negligible, half

of them are fixed for the deleterious allele ($\frac{u}{u+v}$ for asymmetric mutation).

The same reasoning implies that changes in population size will have a negligible effect on the total load in this regime (Figure 2.11). While changes in population size affect the proportion of segregating sites and thus their contribution to load, so long as the population mutation rate remains negligibly small ($\beta \ll 1$), the segregating load will remain negligible compared to the fixed load. In the bottleneck model, the proportion of segregating sites decreases to a new equilibrium after the reduction in population size (Figure 2.11A). This explains the decrease in segregating load, which is balanced by an increase in fixed load (Figure 2.10). By the same token, in the growth model, the segregating load increases but is balanced by a decrease in fixed load, resulting in a negligible change to the total load (Figure 2.10 and Figure 2.11B). In this case, however, segregating sites are still far from their new equilibrium at present (see the next section).

The weak selection regime

In the weakly selected regime, selection and drift have comparable effects on the dynamics of deleterious alleles. As a result, at equilibrium, even moderate differences in population size can affect the balance between selection and drift. Changes in population size also shift the balance, and are followed by transient changes at fixed and segregating sites until a new equilibrium is attained. To understand these effects, we consider the behavior at equilibrium and the rate at which it is approached. For

this purpose, it is helpful to use the low mutation rate (LMR) approximation in which mutant alleles at a segregating site have a single origin; in other words, we ignore mutations that arise during the sojourn of a mutant allele from the time it arises on a background fixed for the other allele to the time it reaches fixation or loss in the population.

The effect of population size on the proportion of sites fixed for the normal and deleterious alleles. At equilibrium, the rate at which deleterious alleles arise and fix is equal to the rate at which normal alleles arise and fix. This balance can be written as

$$2Nup\pi(-2Ns, h, \frac{1}{2N}) = 2Nvq\pi(2Ns, 1-h, \frac{1}{2N}),$$

where π denotes the fixation probability, which depends on the scaled selection and dominance coefficients and on the initial frequency [76] (because $s \ll 1$, we ignore second order terms in s). For $s \ll 1$ and any dominance coefficient, this yields

$$\frac{q}{p} = \frac{u}{v}\frac{\pi(-2Ns, h, \frac{1}{2N})}{\pi(2Ns, 1-h, \frac{1}{2N})} \approx \frac{u}{v}e^{-2Ns}.$$

Namely, at equilibrium, the proportion of fixed deleterious sites declines exponentially with the scaled selection coefficient $\alpha = 2Ns$ (Figure 2.12A). Thus, for a given selection coefficient s, the population size has a dramatic effect on the proportion of sites fixed for the deleterious allele, declining from the neutral, mutation-driven, proportions for $s \ll \frac{1}{2N}$ to approximately 0 for $s \gg \frac{1}{2N}$.

Importantly, however, when the population size changes, the new equilibrium proportion may be attained very slowly. The fractions, $p(t)$ and $q(t)$, of sites fixed for the normal and deleterious alleles t generations after a change in population size (assuming $p(t) + q(t) = 1$) are well approximated by the model

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} p \\ q \end{pmatrix} = \begin{pmatrix} -2N_a u\pi(-2N_a s, h, \frac{1}{2N_a}) & 2N_a v\pi(2N_a s, 1-h, \frac{1}{2N_a}) \\ 2N_a u\pi(-2N_a s, h, \frac{1}{2N_a}) & -2N_a v\pi(2N_a s, 1-h, \frac{1}{2N_a}) \end{pmatrix}\begin{pmatrix} p \\ q \end{pmatrix},$$

where N_a is the population size after the change, and fixation times (on the order of $4N_a$ generations) are neglected. An additional contribution from sites that were segregating before the change is considered below. In this approximation, the change in the fraction of sites fixed for the deleterious alleles is

$$q(t) = q_a^{eq}\left(1 - e^{-\frac{t}{\tau}}\right) + q_b^{eq} e^{-\frac{t}{\tau}},$$

where q_b^{eq} and q_a^{eq} are the equilibrium fractions corresponding to the population sizes before and after the change, and

$$\tau = \left[2N_a\left(u\pi(-2N_a s, h, \frac{1}{2N_a}) + v\pi(2N_a s, 1-h, \frac{1}{2N_a})\right)\right]^{-1}$$

is the timescale of the exponential approach to the new equilibrium. For the semi-dominant case and $s \ll 1$, this time scale is well approximated by

$$\tau \approx \left[u\frac{\alpha}{e^{\alpha}-1} + v\frac{\alpha}{1-e^{-\alpha}}\right]^{-1},$$

demonstrating that it is mutation-limited. This is also true for other dominance coefficients. In other words, following an instantaneous change in population size, the proportion of sites fixed for the deleterious allele will change extremely slowly, at a rate that is inversely proportional to the mutation rate (Figure 2.12B and C).

Because the equilibrium is reached slowly, recent demographic changes in humans should have had little effect on the proportion of sites fixed for the deleterious alleles and hence on the fixed load. The bottleneck at the Out-of-Africa split is estimated to have reduced the population size from $\sim 14,000$ to $1,800$ approximately 2000 generations ago [215]. Once a new equilibrium is reached, there will be a substantial increase in the proportion of fixed deleterious alleles; for example, for a semi-dominant deleterious allele with selection coefficient of $s = 10^{-4}$, it would increase it from 0.05 to 0.4. Yet the change over 2000 generations is minimal, increasing this proportion only by $3 \cdot 10^{-5}$. The estimated 200 generations since the onset of rapid growth in humans is similarly much too short a time period for any measurable effect on the fixed load (which in this case would decrease over large time periods).

The effects of population size on segregating sites. First we consider how the equilibrium properties of segregating sites depend on population size in models with constant population size (Figure 2.13). The deleterious allele frequency at segregating sites decreases with increasing population size, because the efficacy of selection is greater in larger populations (Figure 2.13A). In turn, the proportion of segregating sites increases with population size due to the (linear) increase in the number of mutations that enter the population every generation (Figure 2.13B).

This is true not only for the population as a whole but also for subsamples from it of any size (Figure 2.13C). Finally, the deleterious allele frequency and proportion of segregating sites decrease with increasing dominance coefficient, as stronger selection in heterozygotes results in stronger selection on deleterious mutations (regardless of their frequency) and thus in a shorter sojourn through the population. Thus, in larger populations or if the dominance coefficient is greater, we expect a greater proportion of segregating sites with deleterious alleles at lower frequency.

The total load decreases monotonically when the population size increases (as can be shown using the stationary distribution based on the diffusion approximation [53], for example). This is not true of the segregating load, because the increase in the mutational input can have a greater effect than the increase in the efficacy of selection (Figure 2.13D). Indeed, for selection coefficients closer to neutrality, the increase in mutational input (and the proportion of segregating sites) dominates, causing the segregating load to increase with population size (akin to the behavior in the effectively neutral regime). In contrast, for selection coefficients closer to the strong selection regime, the increase in the efficacy of selection dominates, leading to a reduction in segregating load (akin to the stronger selection regime; see section 2.8.2).

Next we consider the effects of a change in population size. We begin by noting that, for a given population size, the expected sojourn time of deleterious and beneficial mutations that reach fixation is shorter than that for a neutral mutation and is thus on the order of $4N$ generations or less [53]. This implies that on the order of $4N_a$ generations after a change in population size, most of the *old mutations* (i.e.,

those that segregated before the population size changed) have been absorbed (either due to loss or fixation), and replenished by *new mutations* (that arose and spread through the population at its new size). When this turnover process is complete, new segregating sites approach their equilibrium proportions (given a background of fixed sites).

In the bottleneck model, the reduction in the efficacy of selection causes an increase in total load, where the behavior of the components of load can be understood as follows (Figure 2.14). Focusing first on the contribution of old mutations to the fixed load: When old mutations are absorbed, the reduction in the efficacy of selection leads more deleterious alleles to fix than would have had the population size remained constant (at the larger size), eventually resulting in an increase in fixed load. The increase can be approximated by

$$\Delta(s,h,u,N_b,N_a) = \int_0^1 \left(\pi(-2N_a s,h,x) - \pi(-2N_b s,h,x)\right) f(x;h,2N_b s,2N_b u)dx,$$

where $f(x;h,2N_b s,2N_b u)$ is the stationary distribution before the change in population size [53]. The increase is maximized for selection coefficients at which the change in population size leads selection to transit from strong to weak, and is negligible outside this range (Figure 2.14A; explaining why it is more pronounced in Figure 2.14C and D than in E and F, correspondingly). The increase in deleterious fixations and load is then followed by a long-term, slower increase in the fixed load due to new mutations (Figure 2.14C-F). In the parameter regime where the fixation of old mutations makes a substantial contribution to load, there is also a

transient increase in segregating load before the mutations fix (in Figure 2.14C for example). These effects are more pronounced in the recessive case, because of the greater frequency and proportion of segregating sites. Now focusing on the segregating load (Figure 2.14B): when segregating sites attain equilibrium, the reduction in population size causes a decrease in segregating load for lower selection coefficients (Figure 2.14C and D) and an increase for higher selection coefficients (Figure 2.14E and F). Thus, for higher selection coefficients in the weak selection range, both old and new mutations contribute to the transient increase in segregating load observed in Figure 2.10. For the lower selection coefficients in this range, the segregating load decreases both in the short and long term but the fixation of old mutations still results in an overall increase to the total load (Figure 2.10). Importantly, however, on the timescale estimated for the bottleneck at the Out-of-Africa split (vertical line in Figure 2.14), these effects amount to a tiny increase in total load (Figure 2.10).

What about in the case of growth? Human population growth is thought to have started a couple hundred of generations ago, ending with an effective population size in the hundreds of thousands and starting from a size that was thirty-fold smaller [215]. Given the estimated growth parameters, there was insufficient time for the deleterious alleles that segregated before the onset of growth to change their frequencies substantially. Indeed even with the increase in the efficacy of selection as the population size increases, in this regime, selection is too weak to have caused a substantial change in allele frequency over hundreds of generations (although it could have caused the absorption of very rare or very high frequency alleles). After growth, the resulting frequency spectrum of deleterious alleles thus reflects a su-

perposition of the spectrum of segregating sites before growth and of the spectrum at the large number of sites in which mutations were introduced after the onset of growth (Figure 2.15). The many new mutations remain at low frequencies. Because of an increase in the proportion of segregating sites, the segregating load increases at the expense of fixed load, but with negligible effects on the total load, given both the low frequency of new mutations as well as the opposing contributions of normal and deleterious mutations (Figure 2.10).

The strong selection regime

In this regime, purifying selection is sufficiently strong to prevent deleterious alleles from reaching high frequencies, let alone fixation. It follows that there is only segregating load. If we assume that the deleterious allele frequency is small and that the dominance coefficient is sufficiently large, then the load is well approximated by

$$l(h, s) \approx 2hsE(q).$$

Stated another way, when selection against heterozygotes is sufficiently strong, then deleterious homozygotes would be too rare to affect load. Under these assumptions, the diffusion approximation at equilibrium with a constant population size [53] yields

$$E(q) \approx \frac{u}{hs},$$

implying that the load is well approximated by

$$l(h, s) \approx 2u.$$

We refer to the cases where these conditions are met as quasi-dominant.

In the recessive case, the load depends on the second moment of deleterious allele frequency. Assuming once again that the deleterious allele frequency is small, the diffusion approximation at equilibrium with a constant population size [53] yields

$$E(q^2) \approx \frac{u}{s},$$

implying that the load is well approximated by

$$l(0, s) \approx u.$$

The expressions for load in both cases are identical to the classic ones for mutation-selection balance, which are derived assuming an infinite population size [76]. They imply that at equilibrium, the load depends neither on the selection coefficient (explaining the plateaus in Figure 2.9) nor on the population size.

When the dominance coefficient is sufficiently small, however, the load does depend on population size (Figure 2.16). This will be the case when selection against heterozygotes is weak, i.e. when $2Nhs \gg 1$ does not hold, as then both moments of deleterious allele frequency make comparable contributions to load. Holding the

selection coefficient and population size constant, in this range of dominance coefficients, the load varies continuously with h between u and $2u$ (Figure 2.16A). In turn, holding $h \ll 1$ and $N \gg 1$ constant, increasing s also leads the load to vary from u to $2u$ (Figure 2.16B).

Next, we consider the effect of changes in population size, for the quasi-dominant and then the recessive case. We show that in the quasi-dominant case, the load remains constant and is well approximated by the classic derivations for mutation-selection balance. In the recessive case, the load exhibits transient changes before it returns to its equilibrium level.

The quasi-dominant case

In the quasi-dominant case, we can assume deleterious alleles are sufficiently rare that selection against deleterious homozygotes can be ignored and selection has negligible effects on average fitness. Under these conditions, we can approximate the trajectory of a deleterious allele using a branching process (cf. [59]), in which the number of copies that a given deleterious allele gives rise to in the next generation follows a distribution that is independent on the frequency of deleterious alleles in the population.

Consider a single deleterious allele that was introduced by mutation at time $t = 0$ and denote by $Z(t)$ the number of deleterious alleles that it gives rise to at generation t.

The number of mutant alleles in the next generation can then be expressed as

$$Z(t+1) = \sum_{i=1}^{Z(t)} X_i(t),$$

where $X_i(t)$ denotes the number of offspring of the ith allele at time t and $i = 1, \ldots, Z(t)$. We denote the expected number of offspring of a single allele by λ, i.e., $E(X_i(t)) = \lambda$; if we ignore mutations back to the beneficial allele then $\lambda = 1 - hs$ and if we include them then $\lambda = 1 - hs - v$. The expected number of alleles in the next generation is then

$$E(Z(t+1)) = E(\sum_{i=1}^{Z(t)} X_i(t)) = \sum_{j=1}^{\infty} Pr(Z(t) = j) j E(X_i(t)) = E(Z(t))\lambda, \qquad (2.6)$$

or

$$E(Z(t)) = \lambda^{\tau}. \qquad (2.7)$$

Now consider the expected number of deleterious alleles at mutation-selection balance. For this purpose, we measure time backwards from the present. We denote by $Y_\tau(\tau)$ the number of mutations introduced τ generations ago and by $Y_\tau(t)$ the number of alleles that they give rise to at time t. The number of deleterious alleles at the present can then be expressed as the sum of contributions from all the mutations in the past, i.e. $\sum_{\tau=1}^{\infty} Y_\tau(0)$, where, from Equation 2.7,

$$E(Y_\tau(0)) = Y_\tau(\tau)\lambda^{\tau}.$$

In turn, the expected number of new mutations in a given generation is well approximated by

$$E(Y_\tau(\tau)) = 2Nu.$$

It follows that the expected deleterious allele frequency is

$$E(q) = \frac{1}{2N} E(\sum_{\tau=1}^{\infty} Y_\tau(0)) = \frac{1}{2N} \sum_{\tau=1}^{\infty} E(Y_\tau(\tau))\lambda^\tau = \frac{u}{hs},$$

and thus the expected contribution to load is $2u$ - well-known results for mutation-selection balance.

Next, we consider a changing population size. We denote by $N(t)$ the population size t generations in the past and by $a(t) = \frac{N(t-1)}{N(t)}$ the proportional change in one generation. Now the expected number of new mutations introduced at a given time is proportional to the population size

$$E(Y_\tau(\tau)) = 2N(\tau)u,$$

but the fraction of new mutations in the population remains constant (u). Similarly, the expected number of alleles in the next generation is affected by changes in population size

$$E(Y_\tau(t-1)) = \lambda a(t) E(Y_\tau(t)),$$

but their fraction is not, because their increase in number is precisely offset by the

increase in population size

$$E(\frac{Y_\tau(t-1)}{2N(t-1)}) = \lambda a(t)\frac{N(t)}{N(t-1)}E(\frac{Y_\tau(t)}{2N(t)}) = \lambda E(\frac{Y_\tau(t)}{2N(t)}).$$

It follows that the proportional contribution of alleles to the present is the same as that in a constant population size:

$$E(\frac{Y_\tau(0)}{2N(0)}) = u\lambda^\tau,$$

leaving the deleterious allele frequency and the load at the present unchanged (at $\frac{u}{hs}$ and $2u$). In other words, the expected frequency of deleterious alleles and therefore the load follow the same deterministic dynamic as they do in a population of constant size, because when the population size changes, the increase (decrease) in the copy number is precisely offset by the increase (decrease) in population size.

We note that incorporating reverse mutation and migration will not change this conclusion. Reverse mutation would reduce λ, while introducing migration would be similar to both decreasing λ (due to migration of deleterious alleles out of the population) and increasing the mutational input (due to migration of deleterious mutations into the population).

Our results clarify how the expected deleterious allele frequency and proportion of segregating sites at equilibrium depend on population size. When the population mutation rate is sufficiently low, a site switches intermittently between having no deleterious alleles and having a single mutation (by origin) in the population (Fig-

ure 2.17A). Under these conditions, in a larger population size, the mutational input is larger and thus the proportion of time that a site is segregating increases (Figure 2.17B). Because the trajectory of a mutation in terms of numbers of copies does not depend on the population size, the frequency of the mutation is proportional to $1/N$, so the expected frequency of deleterious alleles at segregating sites scales with $1/N$ (Figure 2.17C). In turn, when the population mutation rate is sufficiently high, deleterious alleles are almost always present and often have several mutational origins. Under these conditions, the proportion of segregating sites approaches 1 (Figure 2.17B). Given that the expected frequency at segregating sites is $x = \frac{q}{S_{2N}}$, it follows that the allele frequency asymptotes to $q = \frac{u}{hs}$ (Figure 2.17C). In turn, the variance in allele frequency decreases with population size and asymptotes to 0 in the infinite population size limit.

After a change in population size, a new equilibrium is attained much more rapidly in the strong selection regime because of the rapid turnover of deleterious alleles (see Figure 2.18). However, load is unaffected.

Thinking in terms of the branching process helps us to evaluate previous conjectures about the possible effects of human growth on deleterious alleles. For example, Keinan and Clark [112] suggest that "Some degree of genetic risk for complex disease may be due to this recent rapid expansion of rare variants in the human population". It is indeed the case that the expected copy number of deleterious alleles should be greater under exponential growth; specifically, for a population growing at a geometric rate γ per generation, the copy number will change at a geometric rate of $\lambda + \gamma$

per generation, which will result in an increase if $\lambda + \gamma > 1$. Moreover, population growth increases the sojourn time of a deleterious mutation and, when $\lambda + \gamma > 1$, there is a finite probability it would never go extinct [163]. Importantly, however, the expected *frequency* of quasi-dominant deleterious alleles remains constant, so human population growth has no effect on load.

The recessive case

In this case, the load at equilibrium is again insensitive to population size, but the underlying reasons are quite different than in the quasi-dominant case. In the recessive model, a deleterious allele behaves neutrally while at low frequencies. As a result, its sojourn time (i.e., the expected time that it spends at frequency x) is well approximated by that of a neutral allele (Figure 2.19B). When the frequency x reaches $2Nsx^2 \approx 1$, selection on homozygotes for the deleterious alleles kicks in, and the allele should spend little time above this frequency. In the low mutation rate (LMR) approximation, we can therefore approximate the sojourn time of a recessive deleterious allele as

$$\tau(x) \approx \begin{cases} \frac{2(2N-1)}{1-x} & \text{if } 0 \leq x \leq \frac{1}{2N} \\ \frac{2}{x} & \text{if } \frac{1}{2N} \leq x < \frac{1}{\sqrt{2Ns}} \\ 0 & \text{if } \frac{1}{\sqrt{2Ns}} \leq x < 1 \end{cases},$$

where the expressions for $x < 1/\sqrt{2Ns}$ are the sojourn times (in generations) for a neutral allele (Fig 2.19B). In this approximation, the expected contribution of a

deleterious mutation to load is then

$$s\int_0^1 x^2\tau(x)\mathrm{d}x \approx s\int_0^{\frac{1}{\sqrt{2Ns}}} x^2\frac{2}{x}\mathrm{d}x = \frac{1}{2N},$$

and, given that the expected input of new mutations per generation is $2Nu$, the overall expected load is

$$l(0,s) \approx 2Nu\frac{1}{2N} = u.$$

In other words, (in the low mutation limit) for a given population size N, a recessive allele behaves neutrally up to a frequency of $N^{-\frac{1}{2}}$, resulting in an expected contribution to load that is proportional to N^{-1}. In turn, the mutational input is proportional to N, so they exactly offset.

This back of the envelope approximation also provides an intuitive explanation for the way in which the properties of segregating sites at equilibrium depend on population size (Fig 2.19). First, we consider the proportion of segregating sites (Fig 2.19A). When the population size is sufficiently small for the LMR approximation to apply, the proportion of segregating sites can be approximated by the ratio of the sojourn time of a single mutant through the population to the time between appearances of mutations, namely:

$$S_{2N} \approx \frac{\int_0^1 \tau(x)\mathrm{d}x}{\frac{1}{2Nu}} \approx 2Nu(ln(2N/s)+2).$$

In a larger population size and hence with a larger mutational input, mutations of different origin will overlap, resulting in a slower increase in the proportion of seg-

regating sites with population size. When the mutational input becomes sufficiently large, this proportion asymptotes to 1. Next, we consider the frequency of deleterious alleles. In the LMR approximation, the frequency spectrum of segregating sites can be approximated using the neutral sojourn times up to the threshold frequency $\frac{1}{\sqrt{2Ns}}$ (Fig 2.19B), yielding an average frequency of $E(x) \approx \frac{\frac{2}{\sqrt{2Ns}}}{2+ln\frac{2N}{s}}$. As the population size increases, such that mutations of different origins overlap, the decrease in average frequency becomes slower and asymptotes to $E(x) = E(q) = \sqrt{u/s}$ (Fig 2.19C). Lastly, the turnover time of segregating sites for a given population size N is on the order of $2\sqrt{\frac{2N}{s}}$. As it was for other regimes, this is the time scale for the process of equilibration following a change in population size.

We now consider the implications for the bottleneck and growth models. In the bottleneck model, after the reduction in population size, there is an increase in load followed by a decrease back to the equilibrium level (Figure 2.20A). The transient increase in load (blue arrow in Figure 2.20A) is dominated by the contribution of mutations that segregated before the decrease in population size. The proportion of sites that segregated before was greater and their frequencies lower than after the population size reduction, and while these segregating mutations are gradually absorbed, some of them will drift to higher frequencies, generating a transient surge in load (Figure 2.20B). In turn, the newly introduced mutations have yet to reach equilibrium frequencies and, given that the contribution of the lower frequencies to load is much smaller, they contribute negligibly. In the Tennessen et al. model, the time that elapsed since the bottleneck is longer and the segregating sites are therefore closer to the new equilibrium (green arrow in Figure 2.20A). Correspondingly, the relative

contribution of new mutations is greater and their frequency distribution is closer to equilibrium with the new population size, and yet some contribution from the older mutations remains (Figure 2.20C). These considerations also explain why load exceeds above equilibrium levels in the strong selection regime in Figure 2.10.

In the growth scenario, we see the opposite transient effect: the load is reduced before recovering to its equilibrium level (Figure 2.20D). After the growth period, the number of segregating sites is greatly increased, but the new mutations have had little time to drift to higher frequency. As a result, new mutations segregate at very low frequencies and contribute negligibly to load (Figure 2.20E and F). In turn, mutations that segregated before growth have decreased in frequency due to the increased efficacy of purifying selection, and so their contribution to load declines substantially (Figure 2.20E and F). The result is a transient reduction in load (seen in Figure 2.10 as well as in Figure 2.20D).

Models with dominance coefficients other than 0 and $\frac{1}{2}$

Here we provide summaries of simulations with dominance coefficients other than 0 and 1/2 to illustrate that the same qualitative behaviors are observed. As shown in Figure 2.21, all of the observed qualitative behaviors are included in our previous analysis and summarized in Table 2.1, with one possible exception.

The exception is in the bottleneck model in cases with dominance coefficients $h > 1/2$, where the total load is reduced for lower selection coefficients in the weak se-

lection regime. The reason for this reduction in load is analogous to that for the increase in load that we saw in the recessive case in the same selection regime. For dominance coefficients greater than half, the extinction of low frequency deleterious alleles that segregated before the reduction in population size decreases load more than the fixation of high frequency deleterious alleles increases it. The opposite is true for dominance coefficients smaller than half.

2.8.3 Data analysis and interpretation

We used data from Fu et al. (2012) [68] and from the 1000 Genomes Project [217]. Allele frequency estimates from Fu *et al.* are available from the NHLBI GO Exome Variant Server (`http://evs.gs.washington.edu/EVS/`). These provide estimates of the derived allele frequencies at exonic SNVs in European- and African-Americans (EA and AA). Variants with allele frequencies 0 or 1 in both EA and AAs were excluded.

The haploid sample sizes in Fu et al were EA Autosomal: 8596, EA X: 6717, AA Autosomal: 4434, AA X: 3852. Our primary analysis in the main paper (reported in Figure 3) uses the full sample sizes with the autosomal data. For the purpose of Table 2.2 we wished to compare means on the X and autosomes. Since mean allele frequencies of segregating sites are affected by total sample size, we implemented the following subsampling strategy to facilitate direct comparisons between X and autosomes. First, we converted the reported allele frequencies for each site back into

allele counts (i.e., multiplying each reported frequency by the relevant haploid sample size). Next, we randomly subsampled the autosomal EA and AA variants and the X chromosome EA variant allele frequencies down to a sample size of 3852 chromosomes each, in order to match the haploid sample size for the African-American X chromosome. Subsampling was done without replacement, using the hypergeometric sampling function in R. After sub-sampling, variants whose allele frequencies were both either 0 or 1 were once again dropped. Two-sided t-tests were used to test for allele frequency differences between groups.

1000 Genomes Project vcf files (Phase 1 Version 3) were downloaded from the official 1000 Genomes public server (`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/`). YRI and CEU individuals with (at least) exome sequencing coverage were extracted from the original .vcf files (88 YRI individuals and 81 CEU individuals). 7 YRI individuals, chosen at random, were removed to match sample sizes between YRI and CEU. Variants that were fixed for either allele in both populations were removed. Any variant that was not an SNV or did not contain ancestral allele information was also dropped.

A natural measure for comparing the difference in load between two populations is to count the mean number of derived alleles per individual at SNVs segregating within the joint sample. Note that it is essential in these calculations to define SNVs using the joint sample, otherwise sites that are fixed for the derived allele in Population A but not in Population B would lead to the erroneous conclusion that there are more derived alleles in B than in A.

For our analysis, we found that it is convenient to work with the mean derived allele frequency within each functional class. This quantity allows us to compare frequencies directly between classes, and is also conveniently computed from the Fu et al frequency data. These two measures (mean derived frequency and number of derived alleles per individual) are proportional to one another and hence must yield identical conclusions about the relative load in different populations (for a given functional class: DAF multiplied by twice the number of SNVs yields the number of derived alleles per individual, assuming that missing data have been filled in appropriately). Notice also that we are dealing with mean numbers of alleles, and so these measures are unaffected by deviations from HWE or LE which affect the variance in numbers of derived alleles per individual but not the means.

Of course the number of derived alleles is not equivalent to the number of deleterious alleles, as some variants may be neutral; additionally for weakly selected sites there is a small probability at each site that the ancestral allele is deleterious. Nonetheless, the load is expected to be monotonically increasing with the number of derived alleles. As shown in Figure 2.22, we predict that at semidominant sites there should be essentially no difference in mean derived frequency between AAs and EAs, regardless of selection coefficient. At recessive sites we would expect a small increase in mean frequency in AAs at moderately and strongly selected sites. The fact that we do not observe any significant difference in allele frequencies at "probably damaging" sites argues that the majority of these sites are at least partially dominant.

Mean derived allele frequencies were calculated for both populations at autosomal

noncoding, synonymous, and nonsynonymous sites, as well as autosomal nonsynonymous variants belonging to the different functional categories. Standard errors for each category were estimated using the standard deviation in DAF across sites, divided by the square root of the number of sites in that category. For individual-level analyses, we computed the SD in mean number of variants per individual by bootstrapping across sites. The bootstrap analysis accounts for the evolutionary sampling variance in allele frequencies.

The ANNOVAR suite of scripts [235] was used to obtain functional predictions for each SNP from each of four prediction methods: PolyPhen2 [6], SIFT [121], LRT [32] and MutationTaster [191]. Default program settings were used in each case. The functional designations for each program are as follows: PolyPhen2: D (Probably Damaging), P (Possibly Damaging), B (Benign). SIFT: D (Damaging), T (Tolerant), LRT: D (Deleterious), N (Neutral) and U (Unknown). MutationTaster: A (Disease Causing Automatic), D (Disease Causing), P (Polymorphism Automatic) and N (Polymorphism). Coding versus non-coding and synonymous versus non-synonymous designations were also determined using ANNOVAR. (Note that we also tested the SeattleSeq annotations, and found that the overall numbers were similar (though not identical) to those obtained from ANNOVAR; as with ANNOVAR we found no evidence for a difference in DAF between populations.)

We observed that a strong reference bias exists at sites for which the genome reference sequence carries the derived reference allele. This bias has also been observed by David Reich and Shamil Sunyaev (personal communication). All four functional

prediction programs designate a very high proportion of these sites as being likely nonfunctional or benign, even when the reference allele is rare in the population overall. When we condition on the overall population frequency at these sites, we find that a given site is much more likely to be classified as a probably damaging site if the reference genome carries the ancestral allele than if it carries the derived allele (Figure 2.23).

To deal with this bias, we treated the functional designations at sites where the reference allele is derived as unreliable. As an alternative, we binned all SNVs into a series of allele frequency bins (i.e., the bins shown in Figure 2.23). We assumed that when we condition on the population allele frequency in a very large sample (i.e., the Fu et al sample) that the identity of the genome reference allele carries essentially no further information about the likely functional properties of a variant. Thus, within a bin, the fraction of derived-reference SNVs that fall into each functional category can be predicted from the fraction of ancestral-reference SNVs in that functional category. Thus for example, if 20% of the ancestral-reference SNVs in a given bin have functional category X, then we assume that each of the derived-reference SNVs in that bin has a 20% probability of also being in functional category X. The mean frequency of all SNVs in category X is estimated by summing across all ancestral-reference SNVs in category X plus a sum of contributions from all derived-reference SNVs, weighted by the estimated probabilities that each is in X. As shown in Table 2.3, the bias correction makes a substantial difference to the data analysis. Prior to applying the bias correction, the mean frequency in AAs is substantially higher than in EAs (presumably because more than half of the reference genome sequence is of

non-African origin (Supplement of [82], p145)), but the bias correction makes the two frequencies virtually identical as predicted for models with dominance.

We also provide supplementary results in which we made use of a new unpublished version of PolyPhen's PSIC scores that are calculated in a human-independent (i.e., unbiased) manner. (Thanks to Ivan Adzhubey and Shamil Sunyaev for access to these.) These produce results that are very similar to those from our bias-corrected version, in the sense of showing no difference between populations.

2.8.4 The effects of demography on the genetic architecture of disease risk

A great deal of interest focuses on understanding how recent demographic history has affected the genetic architecture of disease and specifically whether the recent explosive growth has increased the contribution of rare variants to disease risk [34, 112, 158, 215]. Here, we use the theory that we developed to elucidate some of these effects. *Note that while in what follows we refer to disease risk, it also applies to any other quantitative trait.*

A model relating allele frequencies to disease susceptibility

We first consider the relationship between selection on individual loci and disease risk. The few models for this relationship differ sharply in their assumptions. At

one extreme, Pritchard [176] assumed that variants that increase disease susceptibility tend to be deleterious, but that otherwise there is no relationship between the strength of selection acting on these loci and the extent to which they increase disease susceptibility. In turn, Eyre-Walker [55] assumed a correlation between the strength of selection at a locus and its contribution to disease susceptibility. All else being equal, a stronger relationship between the disease risk and fitness implies that the variants that contribute more to disease risk are under stronger selection and, as a result, tend to be younger and rarer. It also follows that their frequency distribution would be more susceptible to the effects of recent demographic events. Here we consider models for the two extremes: one in which the effect sizes are independent on the selection coefficients and the other where the effect sizes are proportional to the selection coefficients.

To model how genetic variation relates to disease risk, we consider the L loci that contribute to disease risk and denote the genotype of individual i at these loci by $\mathbf{G_i} = (g_{i,1}, \ldots, g_{i,L})$. We assume that each of the loci is bi-allelic, with a normal (N) and susceptible (S) alleles, and therefore denote the genotype at locus j ($j = 1, \ldots, L$) as $g_{i,j} = NN$, NS, or SS. We then assume that the probability of developing the disease (ignoring life-history details) takes the form

$$P(\mathbf{G}) = F(\sum_{j=1}^{L} \alpha_j(g_j)),$$

where F is a monotonically increasing function with continuous derivatives that takes

values between 0 and 1 and that

$$\alpha_j(g) = \begin{cases} 0 & \text{if } g = NN \\ h_j a_j & \text{if } g = NS \\ a_j & \text{if } g = SS \end{cases},$$

where h_j and a_j denote the dominance coefficient and effect size of the contribution to susceptibility at locus j. Finally, we assume that the effect of each locus is small, such that we can approximate the variance in susceptibility by *the first term in a Taylor expansion, i.e.,*

$$V(P(G)) \approx [F'(\sum_{j=1}^{L} E(\alpha_j(g_j)))]^2 \sum_{j=1}^{L} V(\alpha_j(g_j)), \tag{2.8}$$

where the variances are taken over the population and

$$V(\alpha(g); x, a, h) = a^2 x(1-x)\left[(2h-1)^2 x^2 + (1-4h^2)x + 2h^2\right],$$

where x is the S-allele frequency.

Our model in which the effect sizes are independent on the selection coefficients (and similarly for dominance coefficients) follows directly. For simplicity we assume that the effect sizes and dominant coefficients are constant, as assuming a distribution yields similar results for all the quantities that we consider below. *The variance in disease susceptibility then follows from Eq. 2.8, where the a_j's and h_j's are constant across loci and the distribution of allele frequencies (the x's) is deterimined by the*

(independent) selection and dominance coefficients (for fitness) at these loci.

Next, we consider the model in which the disease itself is the agent of selection. In other words that the fitness cost results entirely from the probability of developing the disease. Denoting the fitness of affected individuals by W_a and of unaffected by W_u, the relationship between fitness, W, and the probability of developing the disease then takes the form

$$W = PW_a + (1 - P)W_u.$$

In turn, in our model, the relationship between genotype and fitness is

$$W(\mathbf{G_i}) = \prod_{j=1}^{L} w_{i,j} \approx \exp\left(-\sum_{j=1}^{L} \alpha_j(g_{i,j})\right),$$

where

$$\alpha_j(g) = \begin{cases} 0 & \text{if } g = NN \\ h_j s_j & \text{if } g = ND \\ s_j & \text{if } g = DD \end{cases},$$

and we assume that $s_j \ll 1$ and therefore use an exponential approximation. Equating our two expressions for fitness leads to the following model for the relationship between disease risk and genotype

$$P(\mathbf{G}) = \frac{W_u - W(\mathbf{G})}{W_u - W_a} = \frac{W_u}{W_u - W_a} - \frac{1}{W_u - W_a} exp\left(-\sum_{j=1}^{L} \alpha_j(g_j)\right).$$

It follows that under this model, the dominance coefficient and effect size for the

contribution to disease risk equal those for fitness (justifying our use of the same notation for the αs in both).

We now return to the contribution of individual loci to disease risk under this model. Assuming that each locus has a small contribution, i.e., that $\alpha_j(g) \ll 1$ (which follows from $s_j \ll 1$) for $j = 1, \ldots, L$, we can approximate the variance in disease risk by

$$V(P) \approx \exp(-2 \sum_{j=1}^{L} E(\alpha_j(g_j))) \sum_{j=1}^{L} V(\alpha_j(g_j)). \tag{2.9}$$

In other words, the contribution of an individual locus to variation in disease risk is proportional to the variance in fitness at that locus. Here, we consider semi-dominant and recessive loci for which the variances are

$$V(x; s, \frac{1}{2}) = \frac{1}{2} s^2 x(1-x) \tag{2.10}$$

and

$$V(x; s, 0) = s^2 x^2 (1 - x^2), \tag{2.11}$$

correspondingly.

Demographic effects on the variance

Figure 2.24 depicts how different allele frequencies at semi-dominant and recessive loci contribute to the variance in disease risk under the Tennessen et al. [215] model (expanding on Figure 4 in the main text). Because we consider only one selection

coefficient at a time, the relationship between effect sizes and selection coefficient has no effect here; however, we do assume that the dominance coefficient for fitness and for disease risk are the same. The graphs can also be interpreted as the proportional contribution of different allele frequencies to the variance in fitness among individuals. To elucidate the effects of recent demographic events, we also show results for the model with a constant population size (equivalent to the one for the African population before the onset of growth) and for a population that experienced the same instantaneous increase in population size as the ancestral African population in the Tennessen et al. model but then remained constant (from $\sim 7,000$ to $\sim 14,500$ around $6,000$ generations ago, cf. Figure 2.5A), which we refer to as the older growth model.

Demographic effects in the semi-dominant case. First, we consider the effectively neutral regime (Figure 2.24A). In the model with constant population size, the proportional contribution is uniform across frequencies, as expected [53]. In the model of older growth, there is an increased contribution of low and high frequency alleles to the variance (as diversity patterns did not have sufficient time to reach equilibrium yet). In the model for Africans, a similar pattern is observed, with a tiny increase in the contribution from rare alleles due to recent growth (amounting to 0.41% of variance in deleterious variants with frequency below 0.1% and 0.4% in variants above 99.9%). In the model for Europeans, the increase due to growth is also negligible (0.61% of variance in variants with frequency below 0.1% and 0.6% in variants above 99.9%). However, the bottleneck leads to an increased contribution of intermediate frequencies at the expense of moderately low and high frequency alleles

(since low and high frequency alleles are quickly lost or fixed after the reduction in population size).

In the weak selection regime (Figure 2.24B), selection leads to a shift towards lower frequencies and thus to an increased contribution to variance of lower frequency alleles. In turn, the effect of older growth is to increase the contribution of high frequencies: the reason being that before the increase in population size, a greater proportion of sites is fixed for the deleterious allele and at such sites, normal mutations lead to high frequency deleterious alleles. The recent growth in the model for Africans further causes a small increase in the contribution of rare alleles (amounting to 1.4% of variance in variants with frequency below 0.1% and 0.07% in variants above 99.9%). In the model for Europeans, this increase is also small (1.9% of variance in variants with frequency below 0.1% and 0.1% in variants above 99.9%), but the bottleneck again has a substantial effect, increasing the contribution of intermediate frequencies at the expense of lower and higher frequencies.

In the strong selection regime, because of the quick turnover of deleterious alleles, the older increase in population size and the bottleneck in Europeans are too far in the past to have had an effect on alleles that are currently segregating (Figure 2.24C). By the same token, in the Tennessen et al. model, alleles segregating at present are young and therefore the recent growth resulted in a decrease in their frequencies (cf. section 2.8.2), substantially increasing the contribution of rare alleles to variance (with $\sim 70\%$ of the variance contributed by alleles at frequency below 0.1%).

Demographic effects in the recessive case. In this case, recent growth has

little effect in all selection regimes. The contribution of low frequency alleles to variance is much smaller because their effect on load or disease risk is manifested only in homozygotes (Figure 2.24D-F). As a result, the increase in the number of rare deleterious alleles caused by recent growth has a negligible effect on their contribution to the variance in disease risk under both the model for Europeans and Africans (amounting to $\sim 10^{-4}\%$ in the neutral regime, $\sim 5 \cdot 10^{-4}\%$ in the weakly selected and $\sim 0.01\%$ in the strongly selected regime, in variants with frequency below 0.1%). In turn, the increase in the number of high frequency alleles (due to normal mutants on a deleterious background) has a higher impact but it is still quite small (amounting to $\sim 1\%$ in the neutral regime and $\sim 0.2\%$ in the weakly selected regime that are due to variants with frequency above 99.9%).

In the weak and strong selection regimes, there is a peak in the contribution to variance at intermediate frequency (Figure 2.24E and F). Moving from low to intermediate frequencies, the contribution to the variance of a mutant allele increases (see Equation 2.11). This increase is halted, however, because at higher frequencies, selection on homozygotes for the deleterious allele kicks in, leading to few alleles at high frequencies. (Specifically, for a constant population size and given a low mutation rate, the frequency spectrum of deleterious alleles is well approximated by $C\frac{e^{-\alpha x^2}}{x}$, where C is a normalizing constant [53], and thus the contribution to variance can be approximated by $De^{-\alpha x^2}x(1-x)^2$, where D is a normalizing constant.) In the model for Africans (and for older growth), this peak is at higher frequencies in the weak selection regime (Figure 2.24E), because the older increase in population size led to relatively more high frequency alleles at present.

The bottleneck in the model for Europeans has a much more pronounced effect, causing a shift toward intermediate allele frequencies and a corresponding shift in the contribution to variance in all selection regimes (Figure 2.24D-F). As opposed to the semi-dominant case, this is also true for the strong selection regime, as recessive deleterious alleles can reach substantial allele frequencies.

Summary. Population growth increases the relative proportion of rare alleles and could therefore be expected to increase their relative contribution to the variance in disease risk. However, because rare alleles contribute less to the variance to begin with, this effect may be relatively small. Assessing the effects of growth on the genetic architecture of disease risk therefore requires quantification. Here, we have shown that, at least based on current estimates of recent growth, the effects on the variance in disease risk are expected to be negligible. The one exception is the case of strongly selected quasi-dominant alleles, which are young and therefore whose frequencies do reflect the recent population size expansion. Interestingly, in this case, while the architecture of disease risk is substantially affected by growth, the expected load (or disease prevalence) remains unchanged, i.e., the same load will be due to many more deleterious alleles that segregate at lower frequencies than had the population not grown.

In contrast to growth, the bottleneck in European populations should have increased the proportion of intermediate frequency deleterious alleles at the expense of low and high frequency ones (with the exception of strongly selected quasi-dominant alleles, because they are so young). In other words, in these populations, there will be only

a small effect on load but a substantial effect on the architecture of disease, with a greater proportion of the variance in disease risk due to intermediate frequency alleles.

The contribution of rare alleles in a mixture model

In reality, we expect that the variants underlying a complex disease will have a variety of selection coefficients and effect sizes rather than a single one. Under a model with such a mixture, the expected contributions of different allele frequencies to the variance in disease risk can be derived as follows. For simplicity, assume that mutations are semi-dominant (so the dominance coefficient is dropped from the notation). At a site with selection coefficient s, the expected contribution to the variance from deleterious alleles below frequency ω is

$$V_\omega(s) = \frac{1}{2} C E(a^2|s) \int_0^\omega f(x;s) x(1-x) dx, \tag{2.12}$$

where $E(a^2|s)$ is the expectation of the effect size squared for sites with selection coefficient s, $f(x;s)$ is the probability of the deleterious allele being at frequency x (here, we do not condition of the allele segregating) and the proportion coefficient C is akin to the first term in Equation 2.8. The overall contribution to variance of a site is $V_1(s)$ and the fraction of that contribution coming from variants below frequency ω is $\Theta_\omega(s) \equiv \frac{V_\omega(s)}{V_1(s)}$. When all sites are considered jointly, denoting the input of mutations with selection coefficient s by $\mu(s)$, the expected proportion of

variance from deleterious alleles below frequency ω is then

$$\Theta_\omega = \frac{\int_s \mu(s) V_1(s) \Theta_\omega(s) ds}{\int_s \mu(s) V_1(s) ds}. \tag{2.13}$$

Examining the terms in Equation 2.13 suggests that the contribution of rare alleles depends strongly on the relationship between effect sizes and selection coefficients. Specifically, the proportional contribution of rare alleles $\Theta_{0.1\%}(s)$ becomes substantial only for strong selection coefficients (Figure 4D in the main text), as shown in section 2.8.4. The behavior of the overall contribution to variance $V_1(s)$, however, depends on the relationship between effect sizes and selection coefficients. If we assume that the effect sizes do not depend on the selection coefficients (or more precisely that $E(a^2|s)$ is constant) then $V_1(s)$ from weakly selected sites is much greater than from strongly selected sites (Figure 4E in the main text) and rare alleles will make an important contribution only if a very large fraction of the mutational input is at strongly selected sites. If we assume the other extreme in which the effect sizes are proportional to the selection coefficient (or more precisely that $E(a^2|s) \propto s^2$, as in the model in section 2.8.4) then $V_1(s)$ strongly increases with the s (Figure 4E in the main text) and rare alleles would make an important contribution unless the fraction of the mutational input at strongly selected sites is very small. In reality, the outcome could be anywhere in between.

As an illustration, we consider a simple model in which we vary the correlation between selection on variants and their effect on a trait. We assume that half of the newly arising mutations have a weak selection coefficient $s_w = 0.0002$ and half

have a strong selection coefficient of $s_s = 0.01$. For strongly selected mutations, the effect size on the trait, a, is chosen to be cs_s with probability $\frac{1}{2}(1+p)$ and cs_w with probability $\frac{1}{2}(1-p)$, where c is a positive constant and $0 \leq p \leq 1$; correspondingly, for weakly selected mutations the effect size is chosen to be cs_w with probability $\frac{1}{2}(1+p)$ and cs_s with probability $\frac{1}{2}(1-p)$. In this model, the marginal distributions of selection coefficients and effect sizes do not depend on p, while the correlation between them is equal to p. To obtain Figure 4F we therefore vary p between 0 and 1.

2.9 Supplementary Figures

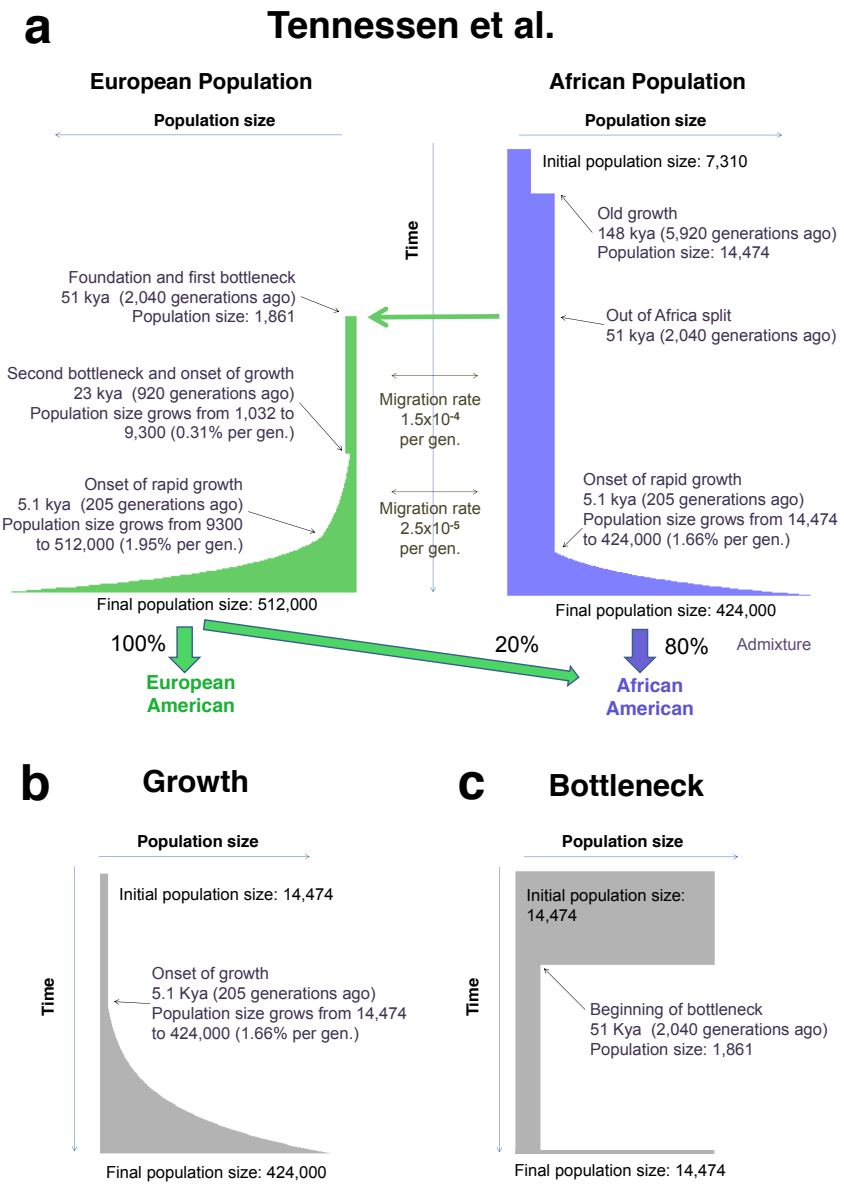

Figure 2.5: The three demographic models that we consider. A) The Out-of-Africa model estimated by Tennessen et al. [215]. C) Exponential growth.B) A population bottleneck. All population sizes are given as number of diploid individuals. In some cases, in order to study the equilibration process, we extend the growth scenario to include a priod with a constant population size after growth and the bottleneck model to include a longer period with a reduced population size.

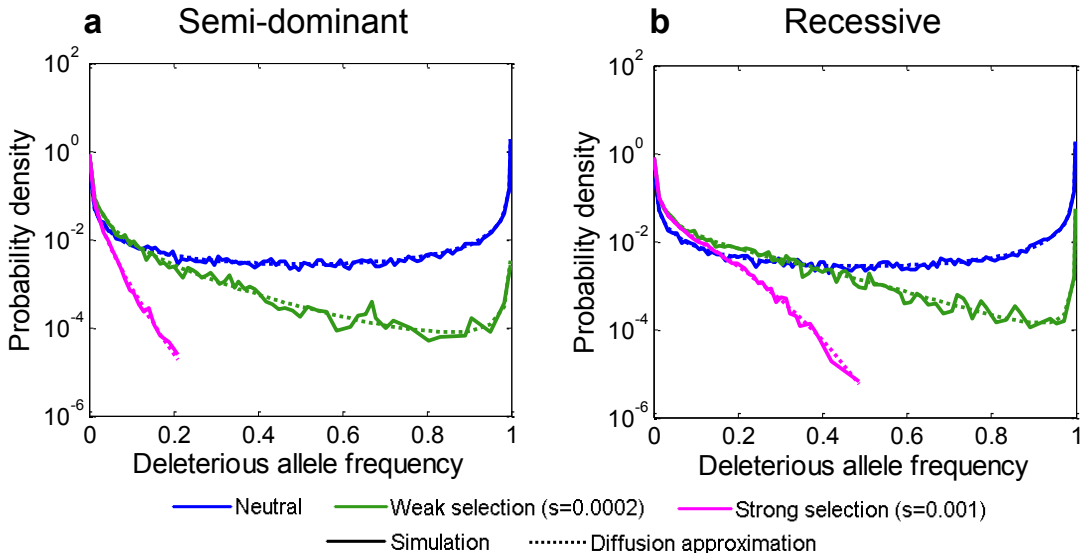

Figure 2.6: Comparison of theoretical and simulated frequency spectra for a constant population size in the (A) semi-dominant and (B) recessive models. Shown are the results based on the diffusion approximation (solid) and on simulations (dashed) for several selection coefficients. The population size was taken as $N = 14,474$ and the mutation rate as $u = 2.36 \cdot 10^{-8}$ per generation per site. The number of runs for each set of parameters was 10^6.

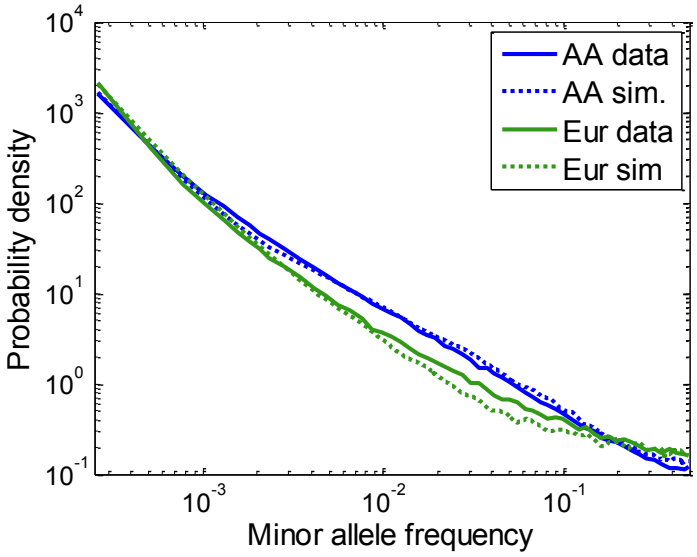

Figure 2.7: Comparison of the minor allele frequency spectrum in data from Fu et. al. and in simulations based on the Tennessen et al. model. The spectra are for a sample size of 3852 chromosomes in AA and EA populations, for both the data and simulations.

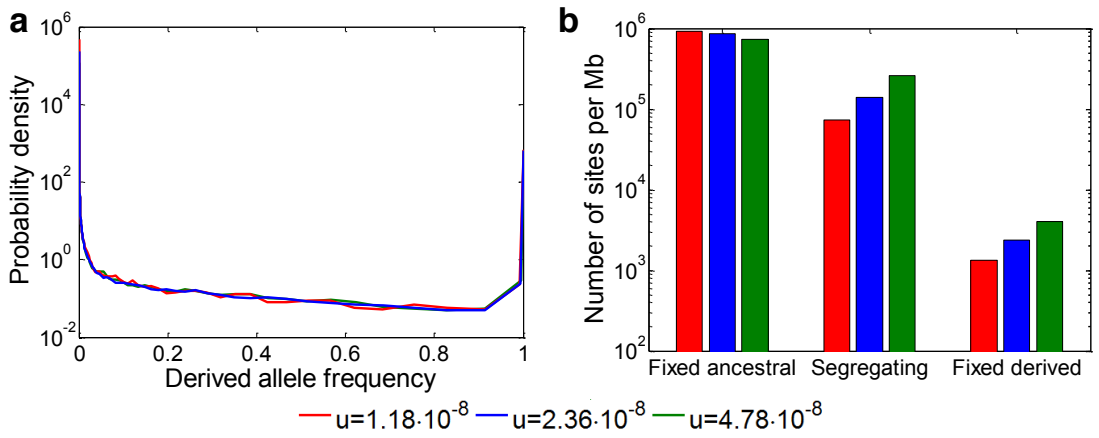

Figure 2.8: Sensitivity of (A) the frequency spectrum and (B) the number of segregating and fixed sites to the mutation rate. The results are shown for simulations of the African population but are qualitatively similar for the European population.

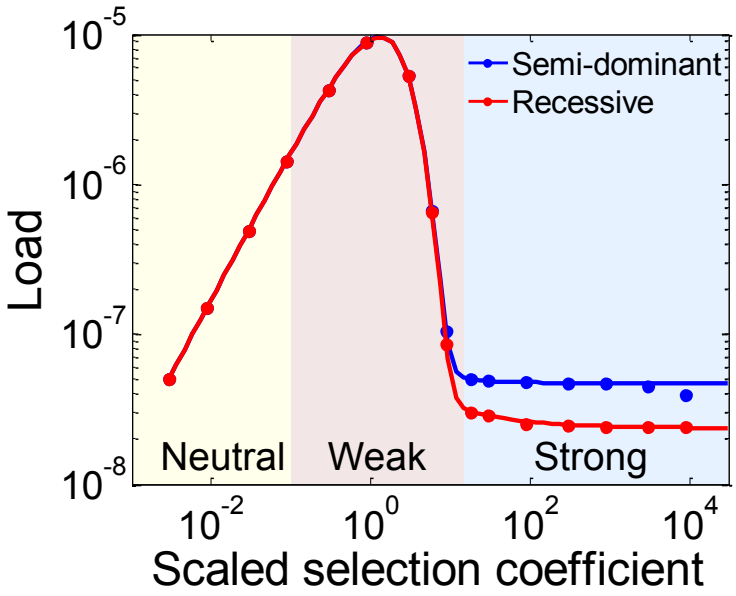

Figure 2.9: Load as a function of selection coefficient in a population of constant size. Results are shown for the semi-dominant (blue) and recessive models (red), where the diffusion approximation is shown as a solid line and simulation results as circles. The population size is $N = 14,474$.

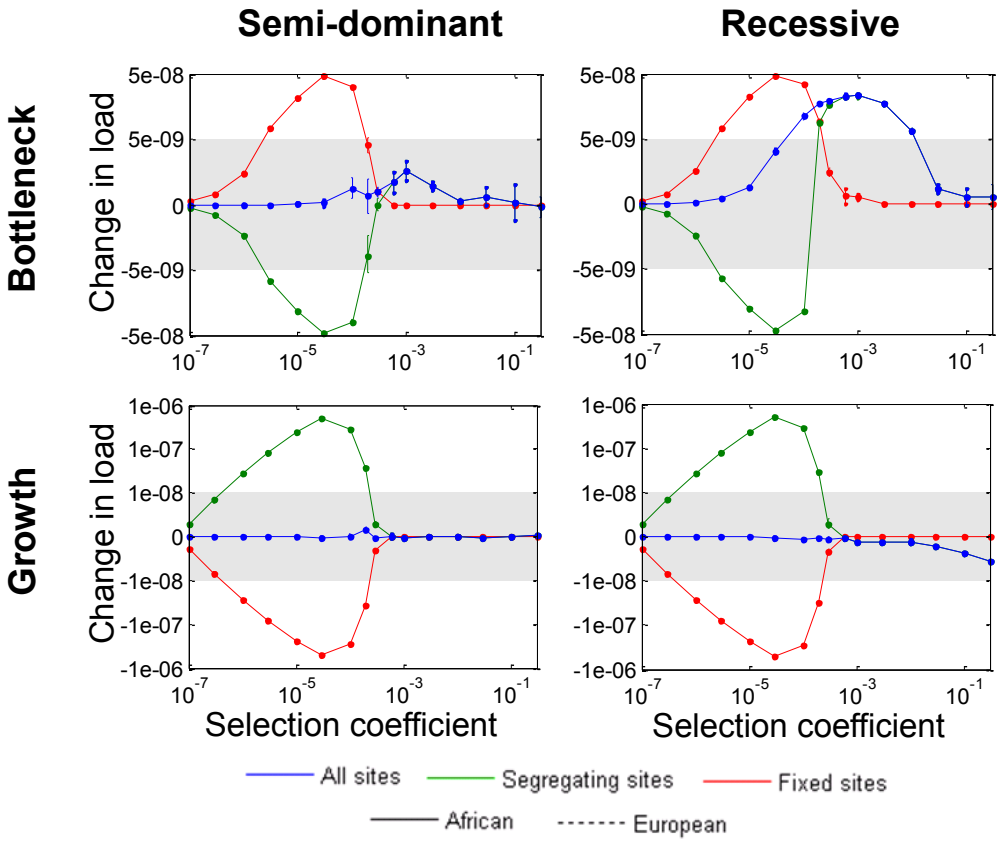

Figure 2.10: The changes to the segregating, fixed and total load under the bottleneck and growth models. Analogous graphs for the Tennessen et al. model are presented in Figure 3 of the main text. Changes are measured by comparison to a population in which the population size has remained constant at the size that it was at the beginning of the demographic model. In the shaded areas, load is shown on linear scale; otherwise it is shown on logarithmic scale.

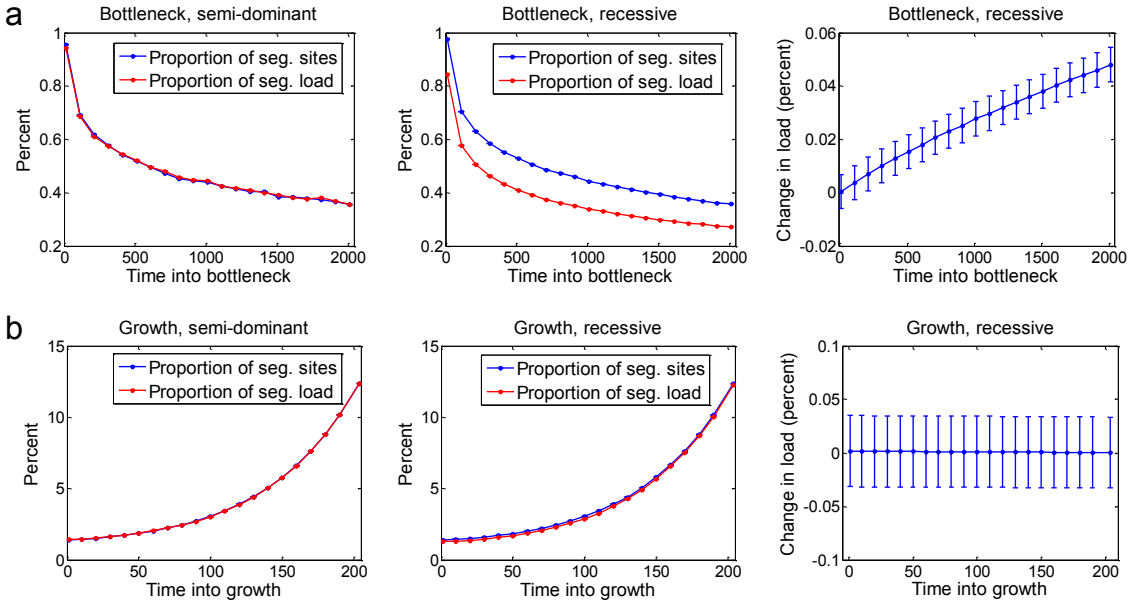

Figure 2.11: Segregating and total load in the bottleneck and growth models in the effectively neutral regime. The proportion of segregating sites, their proportional contribution to load, and the proportional change in total load are shown as a function of time (A) after the bottleneck and (B) since the onset of growth. The selection coefficient is $s = 10^{-7}$. In the semi-dominant case, the expected total load is always $s/2$ regardless of changes in population size; in the recessive case, changes to the proportion of segregating sites affect the total load, but this effect is negligibly small.

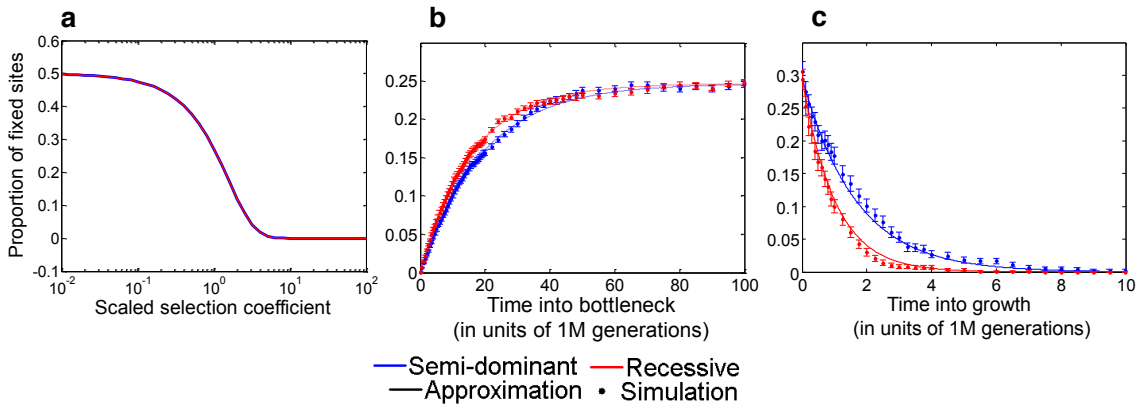

Figure 2.12: Proportion of sites fixed for deleterious alleles in the weak selection regime. In all graphs, the selection coefficient is $s = 10^{-4}$. (A) The equilibrium proportion as a function of the scaled selection coefficient ($\alpha = 2Ns$), where the population size was varied. (B) The proportion as a function of time after the change in population size in the bottleneck model. (C) The proportion as a function of time after the change in population size in the growth model.

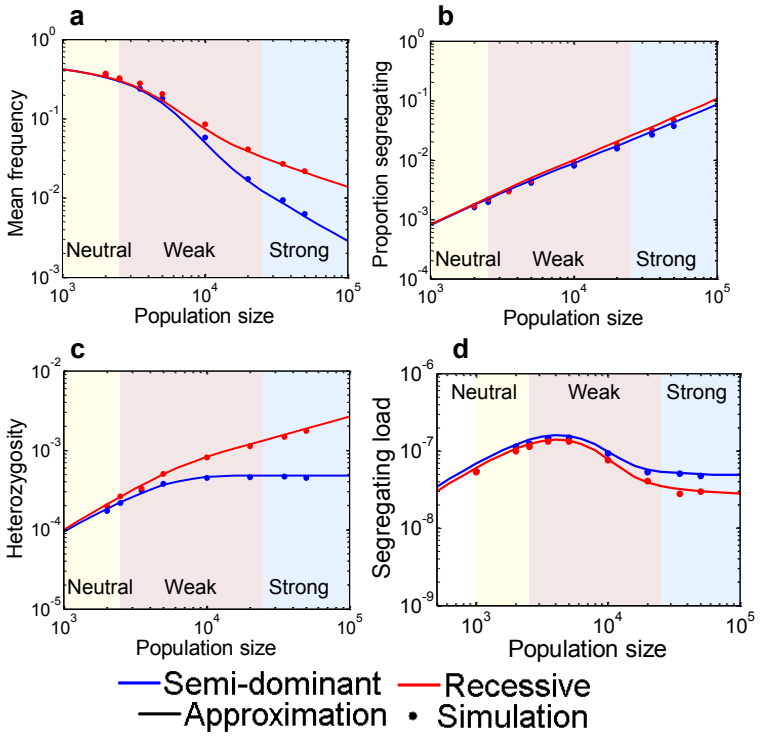

Figure 2.13: Equilibrium properties of segregating sites as a function of population size in constant population size models. In all graphs, $s = 2 \cdot 10^{-4}$. (A) The average frequency of segregating deleterious alleles. (B) The proportion of segregating sites. (C) Heterozygosity. (D) Segregating load.

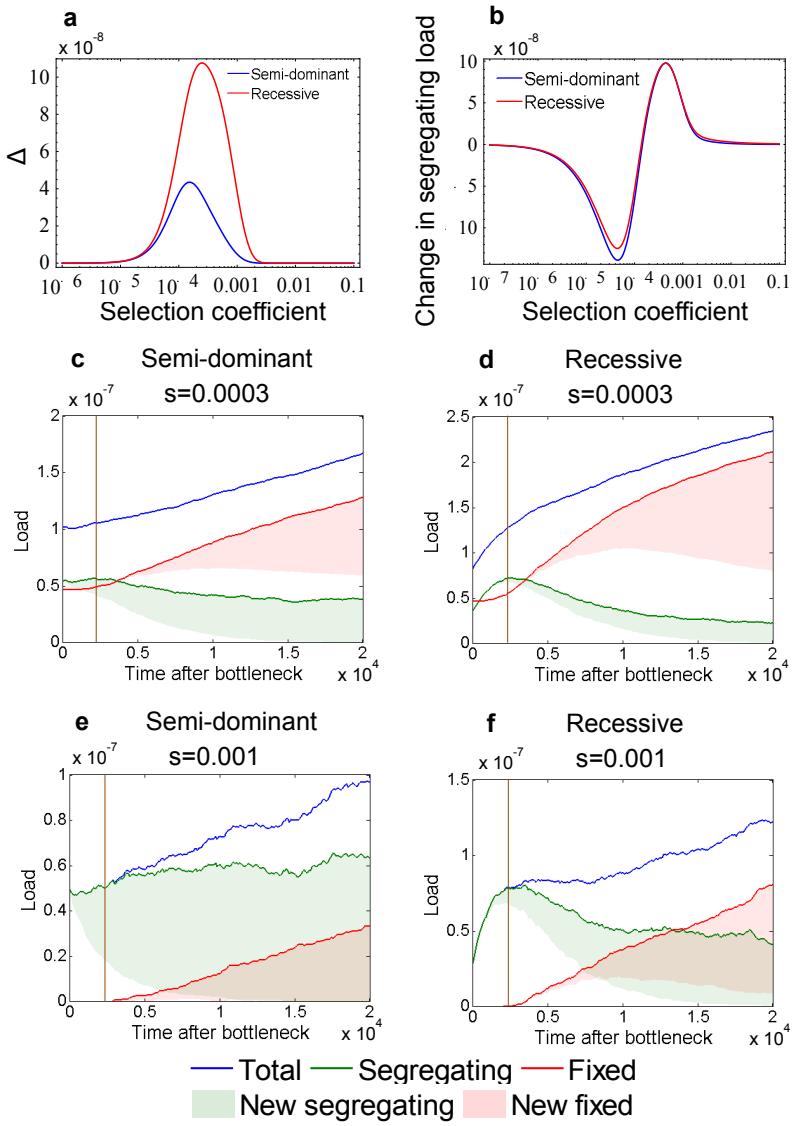

Figure 2.14: The changes in load shortly after a bottleneck. The figure shows (A) the expected change in fixed load due to mutations that segregated before the bottleneck and (B) the expected change in segregating load due to the bottleneck as a function of the selection coefficient. Shown are segregating, fixed and total load from new and all mutations as a function of time since the population size decrease. The semi-dominant (C and E) and recessive cases (D and F) are shown with a selection coefficient in the weak selection regime closer to neutral ($s = 0.0003$) and closer to strong ($s = 0.001$).

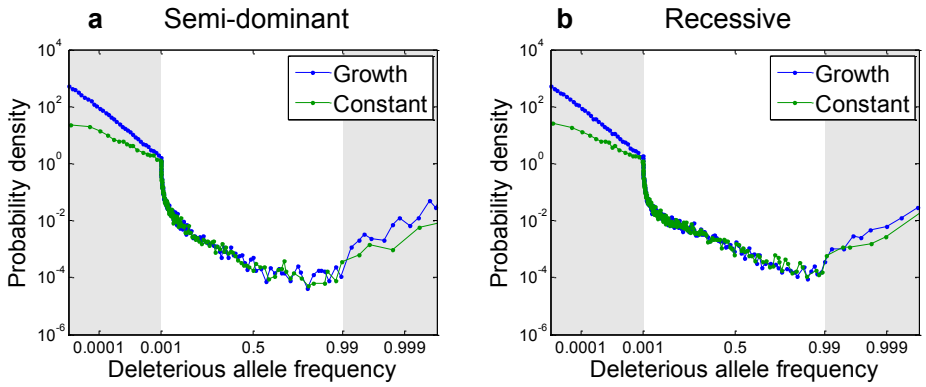

Figure 2.15: The frequency spectrum of weakly deleterious segregating sites in models with and without growth. In the shaded areas, frequency is shown on logarithmic scale; otherwise it is shown on linear scale.

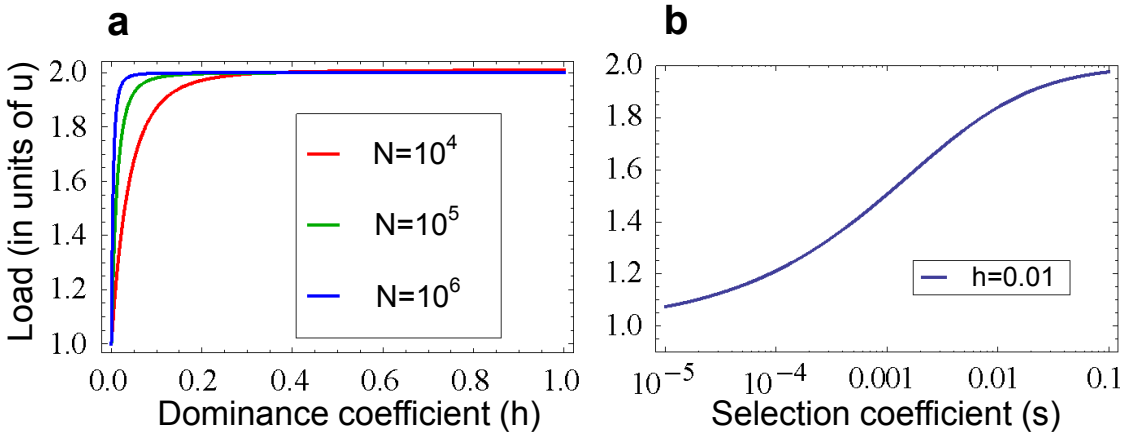

Figure 2.16: The dependence of the load on the dominance coefficient at equilibrium. The graphs were generated using the diffusion approximation for the stationary distribution assuming that the deleterious allele frequency is small [53]. A) Load as a function of the dominance coefficient h, with $s = 0.01$ and population size $N = 10^4, 10^5$ and 10^6. B) Load as a function of the selection coefficient s, with $h = 0.01$ and $N = 10^6$.

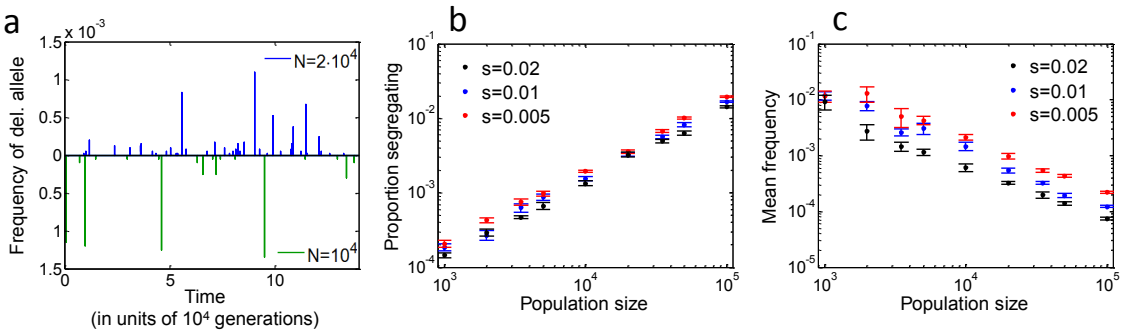

Figure 2.17: The equilibrium properties of segregating sites in the quasi-dominant case. In all graphs, $h = 0.5$ and $u = 10^{-8}$. A) Frequency of deleterious alleles as a function of time in simulations with two population sizes, corresponding to $N = 10^4$ and $2 \cdot 10^4$. In both cases, $s = 0.01$. B) The expected proportion of segregating sites as a function of population size. C) The expected frequency of deleterious alleles at segregating sites as a function of population size.

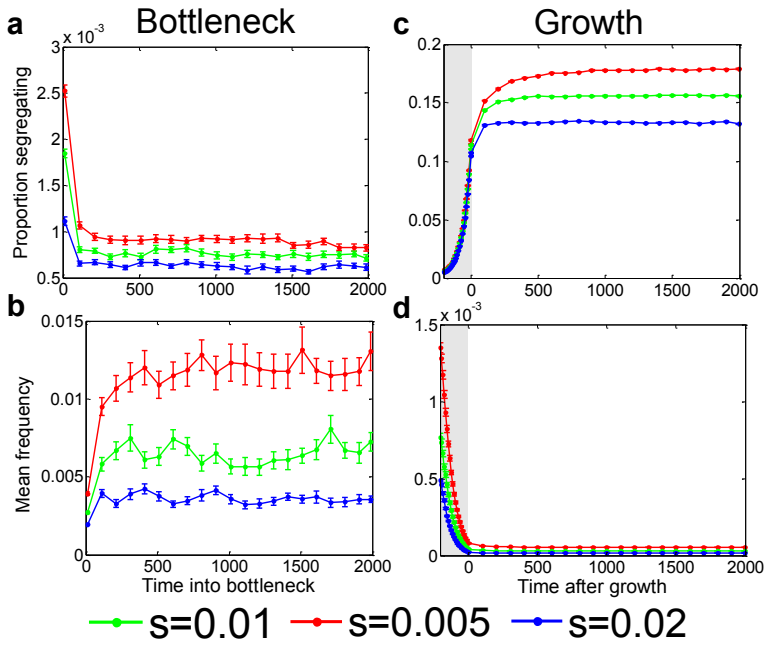

Figure 2.18: The properties of segregating sites as a function of time for the quasi-dominant case. In all graphs, $h = 0.5$. The proportion of segregating sites after (A) the reduction in population size in the bottleneck model and (C) the onset of growth. The expected frequency of deleterious alleles at segregating sites after (B) the reduction in population size in the bottleneck model and (D) after the onset of growth. The shaded region is the period of growth in the Tennessen model.

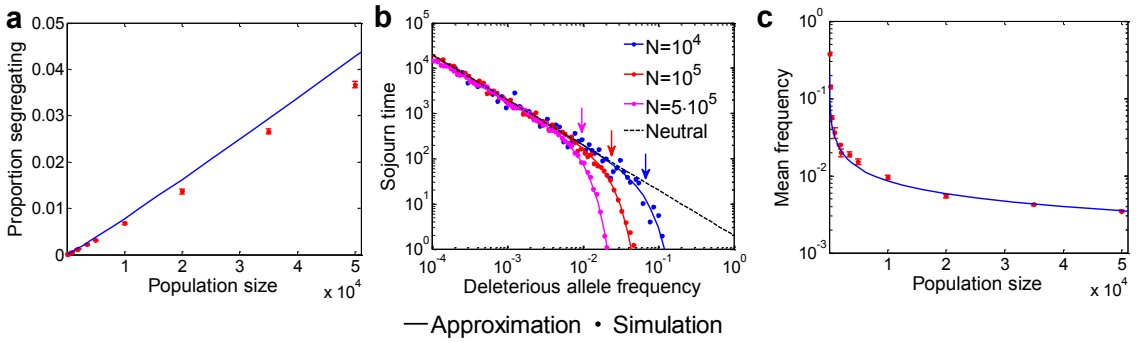

Figure 2.19: The properties of segregating sites at equilibrium in the recessive case, as a function of population size. The selection coefficient is $s = 0.01$. (A) The proportion of segregating sites.(B) The sojourn time of deleterious alleles for different population sizes. The threshold frequency of $\frac{1}{\sqrt{2Ns}}$ for each population size is marked by an arrow with the corresponding color. (C) The average frequency of deleterious alleles.

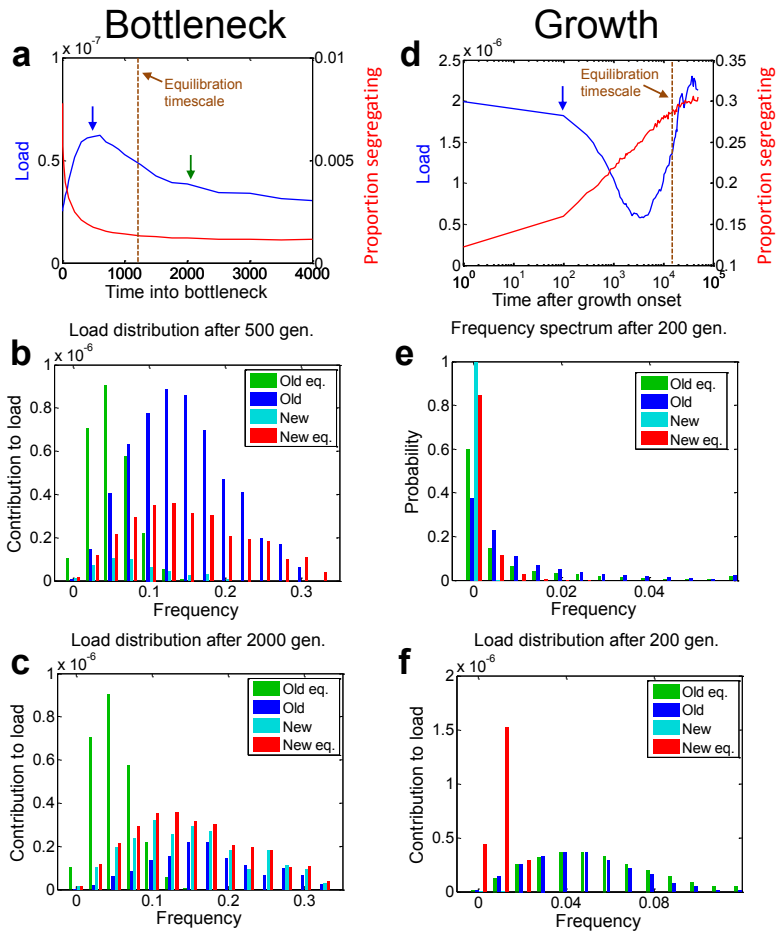

Figure 2.20: Load as a function of time in the recessive case. The selection coefficient is $s = 0.01$. A) The load and proportion of segregating sites as a function of time after the reduction in population size. B) The contribution to load of old and new mutations as a function of frequency, at the time of peak load (500 generations after the reduction in population size, indicated by a blue arrow in A). C) Same as B but for the time since the Out-of-Africa bottleneck, i.e., 50Kya (indicated by a green arrow in A). D) The load and proportion of segregating sites as a function of time after the onset of growth. E) The allele frequency distribution of old and new mutations at the end of the growth period (200 generations after onset, indicated by an arrow in D). F) The contribution to load of old and new mutations as a function of frequency at the end of the growth period.

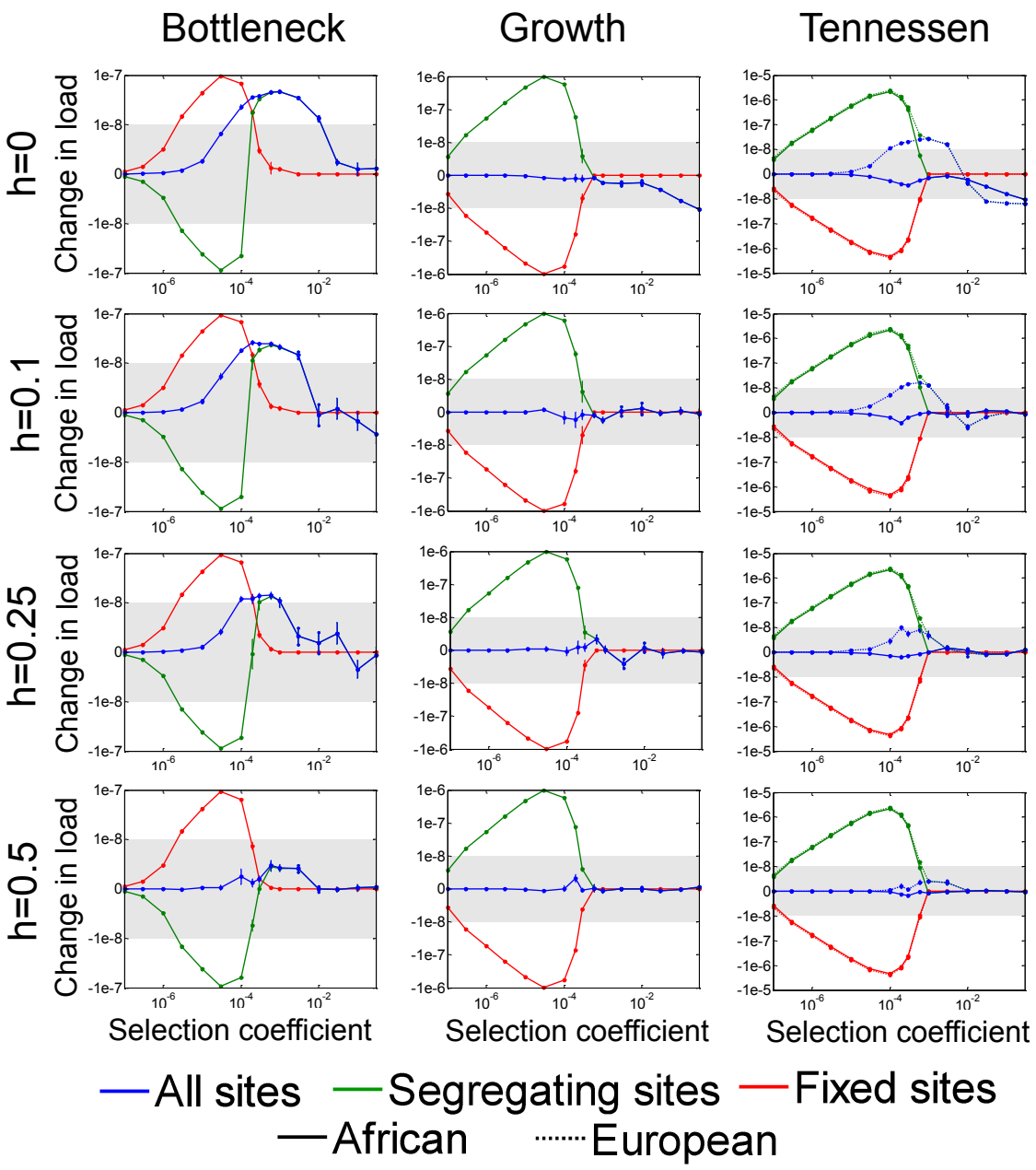

Figure 2.21: Changes in load under the three demographic models with different dominance coefficients. $h = 0$ and $1/2$ correspond to the results in Figure 2.10 and are provided for comparison.

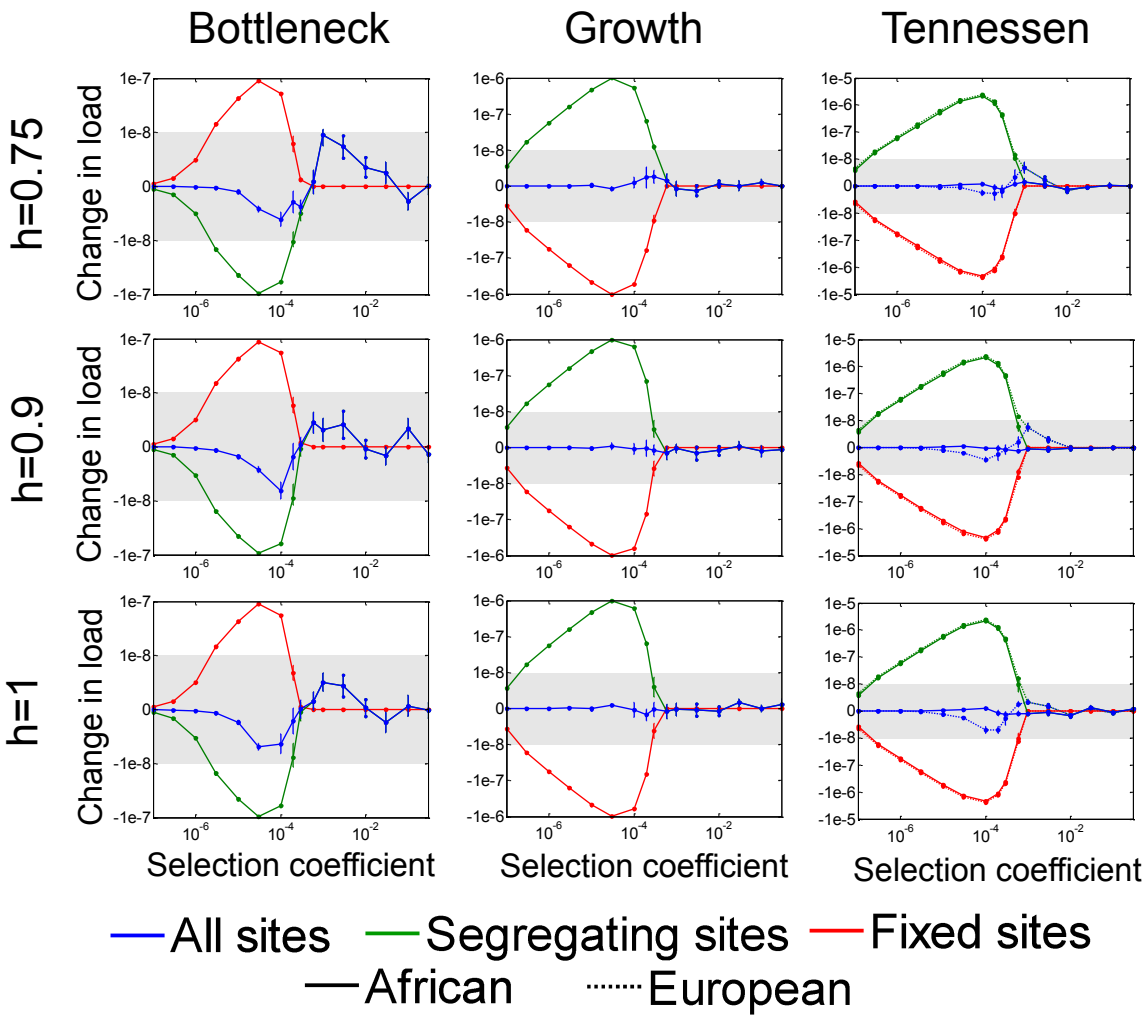

Figure 2.21 (Cont.): Changes in load under the three demographic models with different dominance coefficients. $h = 0$ and $1/2$ correspond to the results in Figure 2.10 and are provided for comparison.

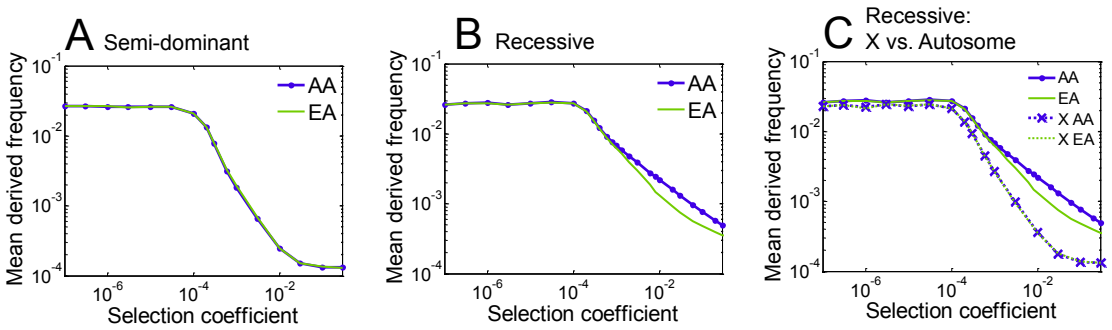

Figure 2.22: Mean derived frequencies predicted as a function of selection coefficient, for the AA and EA demographies. Notice that in (**A**) we predict that for semi-dominant sites AAs and EAs should have essentially identical mean derived frequencies for all levels of selection. In (**B**) we predict a small increase in mean frequencies for AAs at recessive sites with moderate-strong selection. (**C**) provides X vs autosome comparisons under the recessive model; note that recessive alleles on the X experience selection as dominant alleles in males.

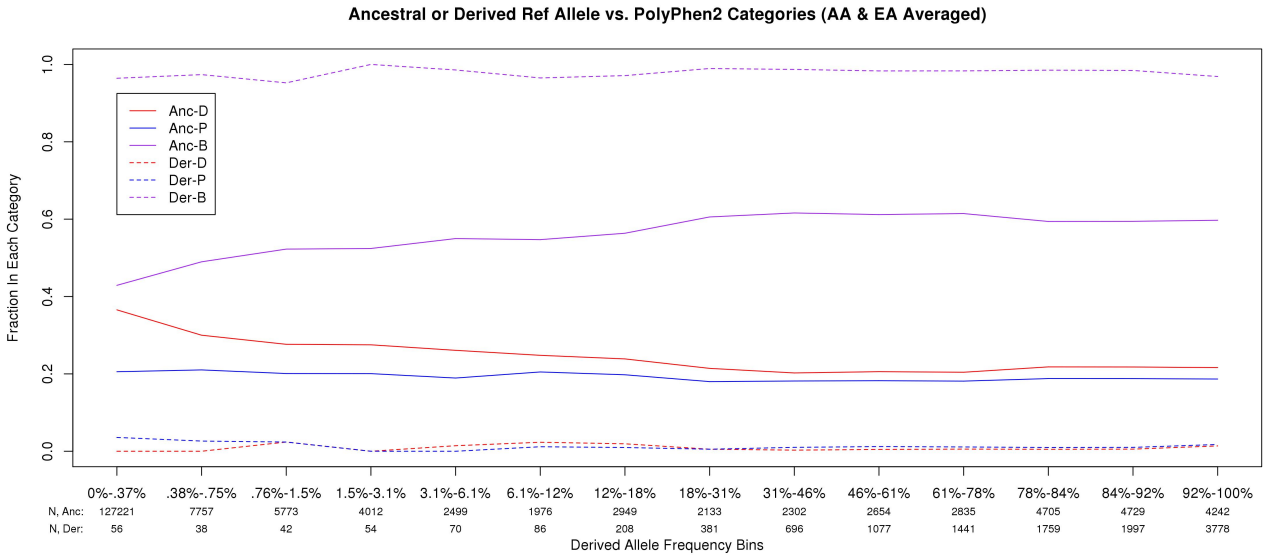

Figure 2.23: Illustration of the reference bias present in PolyPhen 2 [6]. The other functional prediction methods that we considered have a similar bias. The x-axis shows the mean population frequency of nonsynonymous SNVs in the Fu et al data (the left-most bins cover very narrow intervals of frequencies since most of the data are present in these bins). The y-axis plots the fraction of SNVs in each bin that are classified into each of the three PolyPhen categories: **B**enign, **P**ossibly damaging, Probably **D**amaging; and shown separately according to whether the genome reference sequence carries the ancestral or the derived allele. Notice that when the reference carries the ancestral allele, an SNV is classified as Damaging with a probability that ranges from nearly 40% at low frequencies to ≈20% at high frequencies (solid red line). In contrast, for SNVs where the reference carries the derived allele, the fraction of Damaging alleles is near 0% at all frequencies (dotted red line).

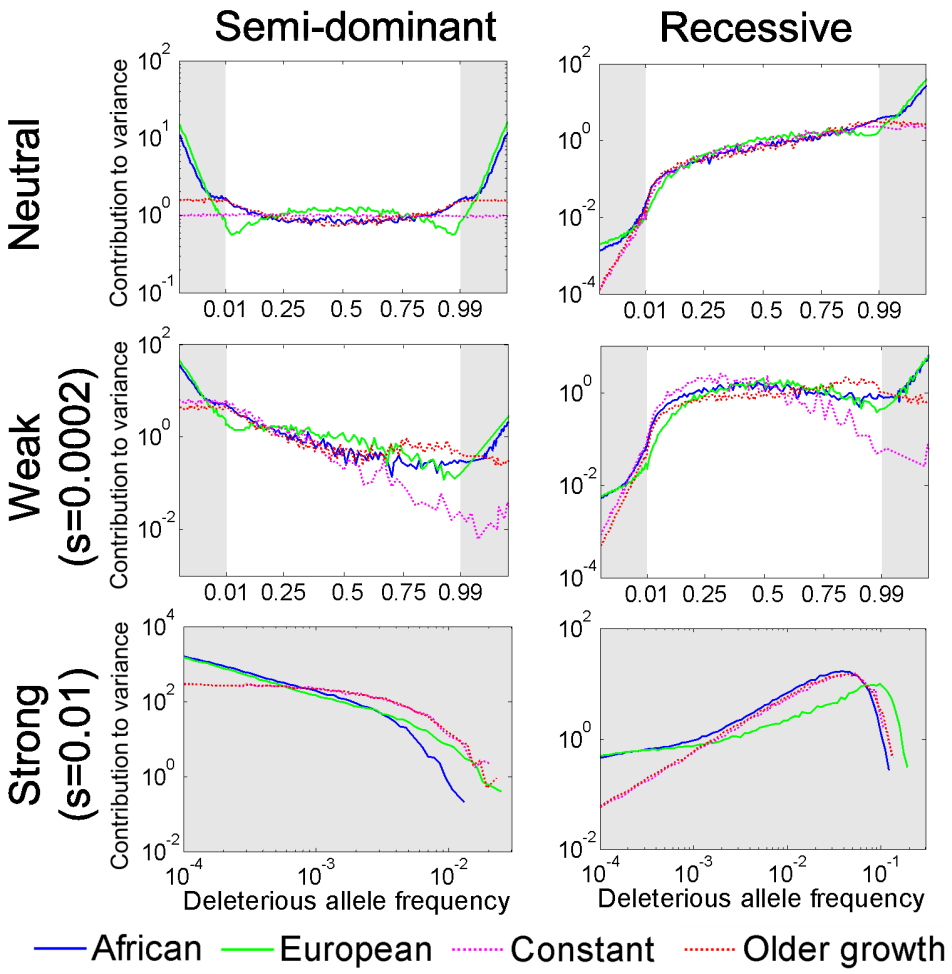

Figure 2.24: The proportional contribution of different allele frequencies to variance in disease risk, under the Tennessen et al. model for Africans and Europeans. Shaded regions correspond to a logarithmic scale on the x-axis, which is included to show the (minor) effects of recent growth.

2.10 Supplementary Tables

			Effectively neutral	Weak		Strong
				closer to neutral	closer to strong	
Bottleneck	Semi-dominant	fixed	increase	increase	increase	—
		segregating	decrease	decrease	increase	unchanged
		total	unchanged	increase	increase	unchanged
	Recessive	fixed	increase	increase	increase	—
		segregating	decrease	decrease	increase	transient increase
		total	unchanged	increase	increase	transient increase
Growth	Semi-dominant	fixed	decrease	decrease		—
		segregating	increase	increase		unchanged
		total	unchanged	unchanged		unchanged
	Recessive	fixed	decrease	decrease		—
		segregating	increase	increase		transient decrease
		total	unchanged	unchanged		transient decrease

Table 2.1: Changes to load under the bottleneck and growth models. The effects on fixed, segregating and total load are depicted by selection regime. The symbol — denotes the cases in which there is no contribution to load both before and after the change in population size.

Method	Chr.	Category	# SNVs	AA_{Mean}	AA_{SE}	EA_{Mean}	EA_{SE}	t-score
Non-coding	Aut	—	300209	0.034	0.00026	0.034	0.00028	0.44
Non-coding	X	—	8355	0.030	0.0015	0.028	0.0016	1.1
Synonymous	Aut	—	220391	0.033	0.00030	0.033	0.00032	0.87
Synonymous	X	—	7001	0.028	0.0016	0.029	0.0018	-0.10
Non-synonymous	Aut	—	351265	0.014	0.00015	0.014	0.00016	0.40
Non-synonymous	X	—	10293	0.012	0.00086	0.012	0.00095	0.076
PolyPhen2	Aut	D	121280	0.0078	0.00011	0.0076	0.00012	1.2
PolyPhen2	Aut	P	65400	0.012	0.00018	0.012	0.00020	0.52
PolyPhen2	Aut	B	132047	0.019	0.00024	0.019	0.00026	0.55
PolyPhen2	X	D	3205	0.0072	0.00065	0.0079	0.00078	-0.99
PolyPhen2	X	P	1957	0.013	0.0012	0.012	0.0012	0.98
PolyPhen2	X	B	3948	0.014	0.0011	0.014	0.0012	0.044
Sift	Aut	D	145986	0.0095	0.00012	0.0093	0.00013	1.6
Sift	Aut	T	180091	0.018	0.00021	0.018	0.00022	-0.13
Sift	X	D	4251	0.0099	0.00076	0.0096	0.00082	0.34
Sift	X	T	5517	0.017	0.0013	0.017	0.0015	-0.29
LRT	Aut	D	146701	0.0060	8.5e-05	0.0060	9.5e-05	-0.11
LRT	Aut	N	160179	0.020	0.00024	0.020	0.00026	0.20
LRT	Aut	U	13845	0.0066	0.00036	0.006	0.00039	2.6
LRT	X	D	3270	0.0038	0.00037	0.0034	0.00034	0.93
LRT	X	N	4548	0.017	0.0014	0.017	0.0016	-0.37
LRT	X	U	886	0.0052	0.0013	0.0046	0.0015	0.40
MutationTaster	Aut	D	155138	0.0022	2.9e-05	0.0017	3.0e-05	18
MutationTaster	Aut	A	5089	0.00089	9.5e-05	0.00056	4.8e-05	4.3
MutationTaster	Aut	N	161169	0.0062	6.8e-05	0.0047	6.7e-05	21
MutationTaster	Aut	P	9040	0.36	0.0047	0.39	0.0051	-6.5
MutationTaster	X	D	3860	0.021	0.0021	0.023	0.0023	-1.2
MutationTaster	X	A	76	0.0010	0.00058	0.00039	0.00017	1.5
MutationTaster	X	N	5566	0.0030	0.00026	0.0013	0.00022	7.0
MutationTaster	X	P	131	0.16	0.028	0.16	0.029	0.28

Table 2.2: Comparison of mean frequencies in AAs and EAs at different classes of sites, classified according to whether the sites are on the autosomes or X, and using a variety of different functional classifications (after application of our bias-correction method). For this table, the data were subsampled down to 3852 chromosomes for AAs and EAs each, to enable X vs autosome comparisons. Note that the mean frequencies in each row are not significantly different ($|t - score| < 2$, with the sole exception of the functional classifications from MutationTaster (which are highly significant). The unusual results for MutationTaster likely arise because MutationTaster uses previously estimated population frequencies in its classification, thus introducing further biases for population genetic analysis that are not properly addressed by correction method.

Method	Chr.	Category	Without bias correction				With bias correction			
			AA_{Mean}	AA_{SE}	EA_{Mean}	EA_{SE}	AA_{Mean}	AA_{SE}	EA_{Mean}	EA_{SE}
Non-synonymous	Aut	—	0.014	0.00015	0.014	0.000162	0.014	0.00015	0.014	0.00016
PolyPhen2	Aut	D	0.0038	9.3E-05	0.0033	1.0E-04	0.0078	0.00011	0.0076	0.00012
PolyPhen2	Aut	P	0.0060	0.00017	0.0053	0.00019	0.012	0.00018	0.012	0.00020
PolyPhen2	Aut	B	0.026	0.00035	0.026	0.00037	0.019	0.00024	0.019	0.00026
Sift	Aut	D	0.0061	0.00013	0.0055	0.00014	0.0095	0.00012	0.0093	0.00013
Sift	Aut	T	0.020	0.00026	0.021	0.00028	0.018	0.00021	0.018	0.00022
LRT	Aut	D	0.0028	6.4E-05	0.0025	7.4E-05	0.0060	8.5e-05	0.0060	9.5e-05
LRT	Aut	N	0.023	0.00029	0.023	0.00031	0.020	0.00024	0.020	0.00026
LRT	Aut	U	0.0081	0.00048	0.0071	5.0E-04	0.0066	0.00036	0.006	0.00039
MutationTaster	Aut	D	0.0017	4.3E-05	0.0011	4.3E-05	0.0022	2.9e-05	0.0017	3.0e-05
MutationTaster	Aut	A	0.0013	0.00034	0.00099	0.00032	0.00089	9.5e-05	0.00056	4.8e-05
MutationTaster	Aut	N	0.013	0.00024	0.012	0.00025	0.0062	6.8e-05	0.0047	6.7e-05
MutationTaster	Aut	P	0.26	0.0027	0.30	0.0032	0.36	0.0047	0.39	0.0051

Table 2.3: Comparison of estimated mean frequencies in samples of 3852 chromosomes, with and without bias correction of the functional annotations. Recall that we observed that all four functional prediction methods typically have low probabilities of assigned 'damaging' status to SNVs where the genome reference carries the derived allele. Notice that prior to applying the bias correction (using all SNVs), AAs tend to have higher allele frequencies at putatively damaging sites, as reported by Tennessen et al. This is likely because most of the reference genome is of non-African origin. After applying our bias correction, we observe that AAs and EAs have essentially identical allele frequencies in all functional categories (except for MutationTaster, likely for reasons discussed above).

Category	AA_{Mean}	AA_{SE}	EA_{Mean}	EA_{SE}	T-Stat
Uncorrected (biased) PolyPhen Scores					
Prob. Damaging	0.00277	6.79e-05	0.00239	7.31e-05	5.4
Poss. Damaging	0.00452	0.00013	0.00401	0.00014	3.84
Benign	0.0208	0.000278	0.0212	0.000297	-1.34
Bias-corrected PolyPhen Scores					
Prob. Damaging	0.00593	8.11e-05	0.00582	8.76e-05	1.23
Poss. Damaging	0.00955	0.00014	0.00948	0.000151	0.488
Benign	0.0154	0.000186	0.0153	2e-04	0.527
Human-independent PolyPhen Scores					
3<PSIC	0.0056	0.0002	0.0054	0.0003	0.45
1.5<PSIC<3	0.011	0.0002	0.011	0.0002	-0.06
PSIC<1.5	0.019	0.0003	0.019	0.0003	-0.07

Table 2.4: Comparison of estimated mean frequencies at autosomal nonsynonymous sites in the Fu et al data, using the full autosomal samples. The top block of data use the uncorrected (biased) PolyPhen scores, and suggest significant differences between populations. The middle block of data applies our bias correction, and shows no significant differences between populations. The bottom block of data uses an unpublished version of the PolyPhen "PSIC" scores that are calculated independent of the human reference sequence, and hence are unbiased (kindly provided by the Shamil Sunyaev lab). These too show no significant difference between populations. Note that DAFs differ between the second two blocks of data due to arbitrary choices in score cutoffs.

Category	YRI_{Mean}	YRI_{SE}	CEU_{Mean}	CEU_{SE}	P-value
Individual-Level Counts					
Synonymous	18,141	119	17,992	122	N.S.
Nonsynonymous	9903	104	9825	80	N.S.
Prob. Damaging	2153	31	2111	26	N.S.
Poss. Damaging	1851	27	1836	24	N.S.
Benign	5899	67	5878	55	N.S.

Table 2.5: Summary of 1000 Genomes Analysis. This table shows the mean numbers of derived alleles per individual in the YRI and CEU populations. The functional categories (Probably/Possibly Damaging and Benign) were obtained from PolyPhen, and adjusted using our bias correction method. SEs obtained by bootstrapping across SNVs. We also obtained identical conclusions (i.e., no difference between populations) when the analysis was done in terms of DAFs, and also when we used the human-independent PolyPhen (PSIC) scores.

CHAPTER 3

IDENTIFYING HUMAN GENES ASSOCIATED WITH HIV-ACQUISITION USING A CANDIDATE GENE-EXOME SEQUENCING APPROACH

Michael C. Turchin[1], Sudhir Penugonda[3], Eun-Yong Kim[3], Kevin Kunstman[3], Matthew Stephens[1,2], and Steven M Wolinsky[3]

[1]Department of Human Genetics and [2]Department of Statistics, The University of Chicago

[3]The Feinberg School of Medicine, Northwestern University, Division of Infectious Diseases, Chicago, IL

3.1 Abstract

Multiple genome-wide association studies (GWAS) have been conducted attempting to link common human genetic variation (minor allele frequency, MAF, >5%) to various aspects of HIV and AIDS pathology. Despite using large sample sizes (up to 10,000s) and samples representing ancestries beyond Western Europeans[57, 168, 151], the majority of associations have only been found within the human leukocyte antigen (HLA) region. Here, we present the results from the first phase of a more focused gene-exome sequencing study looking at ∼1300 genes that have been suggested to interact with the 18 HIV-1 proteins through multiple lines of evidence[22, 117, 104]. We sequenced ∼1300 genes in over 900 individuals from the Multicenter AIDS Cohort Study (MACS)[110] cohort, the majority of which are of Western European descent. We conducted rare-variant (MAF <5%), gene-level tests using SKAT-O and the non-synonymous variants we discovered from our sequencing data. The phenotypes we analyzed included 'HIV-Acqusition' (seronegative individuals vs. seropositive individuals) and 'AIDS-Progression' (slow + very slow progressors vs. rapid + very rapid progressors). Overall, we find a handful of genes outside the HLA region that appear to be marginally associated with HIV-Acquisition. Using these results to prioritize loci, we have begun conducting follow-up genotyping in the full MACS cohort in an attempt to increase our power and further identify molecularly actionable targets involved with HIV-infection.

3.2 Introduction

Human immunodeficiency virus (HIV)-1, a lentivirus that infects and destroys CD4-bearing lymphocytes, was first discovered in the 1980s among gay men in San Francisco and New York[3, 2, 9, 160]. Gradually diminishing and destroying the host's immune-system capabilities, HIV often leads to the development of Acquired Immunodeficiency Syndrome (AIDS), a condition where the host is no longer able to mount proper immune-system responses to foreign agents. As a result, AIDS patients live with a high probability of death due to secondary infections, even from normally-innocuous pathogens, and/or the development of rare cancers[199, 160]. An effective vaccine has yet to be developed, and while there are now anti-retroviral drug therapies available, their effectiveness is not consistent across patients and they generally do not fully eliminate the virus from the patient[9, 85, 160]. Despite HIV/AIDS first being discovered in metropolitan U.S. cities, the virus has now become a worldwide epidemic, with more than 30 million individuals currently living with HIV infection across the globe, 70% of them in the developing world[160, 142, 1].

As researchers began to study this burgeoning outbreak, they quickly noticed natural variation in how humans respond to HIV. Some individuals never become infected despite multiple exposures to HIV, while others once infected never develop AIDS. Furthermore, even among patients that develop AIDS, the time to development can range from just a few years (rapid progessors) to over ten or more years (slow progressors). Researchers posited that this differential response to HIV might be due to the underlying genetic variation present in human populations. As an example,

researchers discovered by the late 1980s that humans with a specific 32-basepair deletion in the gene *CCR5* are immune to HIV-infection[42, 187, 94]. Researchers learned that *CCR5* encodes a T-cell receptor that is necessary for HIV-infection, and that this 32-basepair deletion produces a structural change extensive enough to inhibit HIV from binding. Clearly then, human host-genetics can play a role in mediating response to HIV-infection and AIDS-progression.

Progress in discovering more relevant biology such as CCR5 has been challenging however. During the 1990s, numerous candidate gene studies elucidated possible key factors of the human immune-system involved with HIV/AIDS biology, such as another T-cell receptor CXCR4, the killer immunoglobulin-like receptor (KIR) gene family, and the APOBEC3 proteins[94, 9, 192]. However, during the recent genome-wide association study (GWAS) era of the late 2000s, few of these loci were found to reach genome-wide significance with any HIV/AIDS-related phenotype. Despite using sample sizes exceeding 10,000 and ancestries beyond European-Americans, very few loci outside the human leukocyte antigen (HLA)-region reached genome-wide significance[57, 24, 58, 169, 151]. Encouragingly however, candidate loci would sometimes reach at least marginal levels of statistical significance, suggesting some concordance between previous and current approaches[94, 9, 160, 192]. Overall though GWAS had produced mixed results and host genetics HIV/AIDS researchers, much like other human genetics-related fields, worked to identify immediate future directions. One research avenue that gained traction was to use recently available sequencing technology to target rare variation (variants present at minor allele frequencies (MAF) <5%)[58, 85, 151]. GWAS mainly evaluated com-

mon genetic variants (>5%), leaving this newly accessible part of the allele-frequency spectrum still unexplored. With HIV being a relatively recent disease, it is possible rare variation may play an important role in the underlying genetic architecture of HIV/AIDS-related phenotypes.

Therefore, in this project we take the next steps in elucidating human loci involved with HIV/AIDS and conduct targeted exome-sequencing on a list of human candidate HIV-target genes. Specifically, we present here the first phase of a two-stage analysis. We conduct gene-exome sequencing on a list of candidate loci with strong *a priori* experimental evidence for functionally being related to HIV[22, 117, 104]. We sequence these genes in a subset of the Multicenter AIDS Cohort Study (MACS)[110], a resource enriched for males with varying degrees of exposure to HIV during the 1980s. We then conduct rare-variant, gene-based analyses using the SNPs we discover, focusing on the phenotypes of HIV-Acquisition and AIDS-progression. Using these results, we then develop a list of target variants for the second phase of this study – follow-up genotyping in the full MACS cohort. Follow-up genotyping will emphasize variants from among top candidates as ranked by our first-stage results. Approaches similar to this study design have already produced results in other phenotypes, including LDL cholesterol and colorectal cancer[157, 47, 52, 127], thus suggesting this may be a productive setup. Additionally, this project is being conducted in collaboration with researchers from the HIV Immune Networks Team (HINT); these collaborators have strong expertise in the molecular biology of human-HIV interactions and are readily available to conduct molecular follow-up on our top results.

By using a candidate gene approach we aim to take a focused, hypothesis-based look at the human genome, and by including rare variants in our analyses, we aim to assess the full spectrum of genetic variation present at each locus.

3.3 Results

3.3.1 Sequencing, variant-calling, and QC of MACS subset

Details regarding MACS' subset and candidate-gene selections, as well as targeted 454 gene-exome sequencing, can be found in Methods and Supplementary Table 3.5. In short however, a total of 988 individuals were chosen from the MACS cohort. This subset includes individuals that are either infected with HIV or not (seropositive and seronegative, respectively), individuals that are seronegative but engaged in high-risk sexual behavior (highly-exposed seronegative), and individuals that are seropositive that either slowly progressed towards AIDS or rapidly progressed towards AIDS (very slow/slow progressors and very rapid/rapid progressors, respectively). A total of 1,693 candidate-genes were selected for targeted gene-exome sequencing. Specifically, this gene-exome sequencing was carried out by splitting the candidate list into three custom Nimblegen arrays, which were then processed in-house using 454 GS FLEX+ pyrosequencing. Here we present sequencing, variant-calling, and analytical results from the first two arrays (named MM5 and P2), which total ∼1,300 genes (∼500 and ∼800, respectively) of the original list.

Mapping and variant-calling were predominantly conducted following 1000Genomes[73] and Broad GATK[43] protocols; for details regarding these steps see Methods and Supplementary Figure 3.5. In brief, we mapped sequence reads using the BWA MEM algorithm[134] and we called variants using GATK's Universal Genotyper. At each stage of the pipeline typical quality control (QC) measures were performed to ensure a high-quality dataset, including GATK's base-quality score recalibration and variant-quality score recalibration. QC measures were also performed on the individual level, including removing samples with low genotyping rates and samples that appeared as PCA outliers (see Methods and Supplementary Figure 3.5). Additionally, only white non-Hispanic individuals were ultimately used in the final dataset; the original MACS subset did contain individuals from other ancestries, but their individual numbers were too low to be included as separate analyses (see Supplementary Table 3.5 for MM5 and P2 pre- and post-QC individual breakdowns). The full MACS cohort however contains a much greater number of individuals from each ancestry, so these other subsets will be analyzed in the eventual follow-up genotyping.

These steps produced a final dataset of 775 individuals for MM5 and 747 individuals for P2 being included for analysis. Post-QC, MM5 had an average sequencing depth of 22× per individual and P2 had an average sequencing depth of 14× per individual (see Supplementary Figure 3.8 and Supplementary Table 3.6). Across both arrays a total of 6.0×10^{-8} sequencing reads and 3.6×10^{-11} sequenced basepairs were produced, leading to a final dataset of 149,063 high-quality called variants. These variants include 17,330 exonic and 8,842 non-synonymous SNPs across both arrays, with 473 genes represented post-QC from MM5 and 785 genes represented post-QC from

P2.

3.3.2 External Validation with Public Databases

To help check the performance of our pipeline and QC steps, we compared our finalized variant calls to two external human genomics datasets. Using the European subset of the 1000Genomes[73] dataset and the non-Finnish European subset of the ExAC dataset[131], we directly compared the allele frequencies (AFs) per variant from our dataset to the AFs per variant from these public datasets (see Methods). To accomplish this, we first identified overlapping SNPs between our finalized variant calls and each databases' variant sets; we then plotted the observed AFs of these overlapping SNPs against one another for both external datasets (Figures 3.1 & 3.2 and Supplementary Figures 3.9 & 3.10). In all plots we observe a strong concordance between our variant-called AFs and the databases' AFs, even among rare SNPs (<5% reference AF). For example, compared against the 1000G data, overlapping MM5 variants have a mean AF difference of .13% (s.d. = .59%, # SNPs = 38,859) and overlapping P2 variants have a mean AF difference of .14% (s.d. = 1.3%, # SNPs = 54,910). Across both arrays and both external datasets, these results help suggest that our pipeline produced accurate and high-quality final variant-calls.

3.3.3 SKAT-O Analyses

To test for association between rare variants and our HIV-related phenotypes, we used a 2nd generation rare-variant test SKAT-O[130]. SKAT-O combines the two rare-variant models predominantly evaluated by the first generation of rare variant methods, 'burden-based' test and a 'variance-based'. To explain these two models, imagine we have a case vs. control setup with two groups of individuals differentiated by a phenotype of interest, such as presence and absence of a disease. In a 'burden-based' test we anticipate rare variants for a given gene (e.g. singles, doubletons) to only be present in one set of these individuals, i.e. mutations will **only** ever make you more susceptible or more resistant. Therefore for a test of association we will simply add up the number of rare variants we see as a final test metric. In a 'variance-based' test, we anticipate seeing rare variant for a given gene to be present in either sets of individuals, i.e. mutations can make you **either** more susceptible or resistant. In this scenario, we can no longer just add up variants since they may cancel one another out; instead we measure the variance of the mutation counts across the gene, expecting genes related to our phenotype of interest to have greater values than the background distribution. Ultimately researchers felt that both models may represent true biological scenarios, so 2nd generation methods such as SKAT-O opted for compromises between the two setups[130]. Specifically, SKAT-O accomplishes this by combining both tests as an overall mixture model and systemically testing different proportions of the two scenarios, ranging from 100% 'variance-based' to 100% 'burden-based'.

One important aspect of these rare-variant approaches is that they are generally gene-based; unlike previous GWAS methods that tested each individual variant, here we combine variants to represent loci that are then jointly tested together for association. For SKAT-O, this leads to a design choice: how should we combine variants to define any single gene? We could combine any variant we discovered across the entire locus, or choose a more focused approach by including only non-synonymous or predicted 'damaging' variants. To evaluate these different choices, we took an empirical route and conducted preliminary analyses exploring a handful of possibilities. The design strategies we evaluated included: a) collecting all variants between a 10kb window of the gene's transcription start and end sites (ie exonic, intronic, and intergenic variants) b) only exonic variants c) only non-synonymous variants d) only 'deleterious' variants as determined by Polyphen2[5] and e) only 'deleterious' variants as determined by Sift[120]. Preliminary SKAT-O results using MM5 suggested that power increased as we went from using all variants to just non-synonymous variants, but that power began to plateau or decrease as we further focused on 'deleterious' variants (see Supplementary Figures 3.11-3.14 for early examples from MM5). This later loss in power is likely due to the number of variants being included per test becoming too low. Additionally, by focusing on 'deleterious' variants, some genes were left with no testable SNPs and were subsequently dropped. Therefore, we continued onto our main analyses by using the non-synonymous approach as a compromise between the two extremes of our design strategies.

For our tests of association, we first broadly focused on 2 different phenotypes: 'HIV-Acquisition' and 'AIDS-Progression'. HIV-Acquisition was defined as seronegatives

vs. seropositives, and AIDS-Progression was defined as rapid + very rapid progressors vs. slow + very slow progressors. We then have an additional two phenotypes that represent 'extremes' of these first two setups. For HIV-Acquisition, we have a second phenotype where in lieu of seronegative individuals, we use highly-exposed seronegative individuals; these highly-exposed individuals represent a potentially stricter set of controls. We define this phenotype, highly-exposed seronegatives vs. seropositive, as 'HIV-Acquisition HE'. And for AIDS-Progression, we have a second phenotype where we only use the most extreme progressors for both rapid and slow. We define this phenotype, very rapid progressors vs. very slow progressors, as 'AIDS-Progression Extreme'.

We ran each of these four phenotypes separately on the genes from MM5 and P2. We determined significance by comparing our observed p-values to the distribution of p-values from 1,000 permutations (where phenotypic labels were switched). QQ-plots displaying our observed results and 95% confidence intervals (CIs) derived from our permutations for HIV-Acquisition are shown in Figures 3.3 and 3.4, and QQ-plots from all other analyses (HESN HIV-Acquisition and both AIDS-Progression phenotypes) are shown in Supplementary Figures 3.15-3.20. The strongest top 5 loci found in the HIV-Acquisition analysis for both MM5 and P2 are given in Table 3.1.

For 'HIV-Acquisition', MM5 and P2 both show qualitatively similar results. Comparing our observed p-values to the distribution from permutations, our most strongly associated genes are not outside the boundaries of our 95% CIs. This result however

may be not particularly surprising; with a sample size of under 1,000 individuals we likely only had power to pick up particularly large effect sizes. Going further down into the p-value distribution, we do see an enrichment of loci falling just outside our 95% CIs. This enrichment may be more pronounced in MM5 vs. P2; this is possibly due to the experimental evidence being more direct in the genes underlying MM5 vs. P2. MM5 contains genes from a genome-wide PPI assay whereas P2 contains genes from multiple RNAi screens. It would not be surprising if a PPI study had more accurate findings than an RNAi screen, since there is a large potential for off-target effects with RNAi approaches[103].

For the other phenotypes tested, we see a weaker signal of association in HIV-Acquisition HE, and we see no discernible signal for either of the AIDS-Progression phenotypes. In both these cases, we are likely witnessing the limits of our sample-sizes. Using the highly-exposed seronegative individuals, despite representing a 'more clean' set of controls, still leads to less samples being analyzed (591 individuals in MM5), and both AIDS-Progression analyses contain in total less than 150 individuals (v.rapid + rapid vs. v.slow + slow = 147 and v.rapid vs. v.slow = 63 in MM5).

3.3.4 Pathway Analyses

As a set of complementary analyses to our gene-level association tests, we also conducted pathway analyses. Since we observed no significant results from our gene-level SKAT-O analyses, we were interested in alternative approaches that may increase

our power. One way to increase power for SKAT-O tests is to include more variants per locus analyzed. Therefore, by using pathways in lieu of genes to define the 'loci' we are testing, we are possibly including more variants per analysis; this in turn may potentially increase our power to detect associations.

To adopt our SKAT-O setup for pathway analysis, we make one change to our pipeline. Previously we were defining a 'locus' as all the non-synonymous variants found for a single gene. Now, we will define a 'locus' as all the non-synonymous variants found across all the genes mapped to a given pathway. So for each pathway we will first identify which genes (either from MM5 or P2 separately) are present in our dataset, and then we will collect all the non-synonymous variants among those genes as our testing unit. The actual pathway definitions we will use are derived from two different sources: 1) the Broad Institute's MSigDB C2 pathway collection, a manually curated grouping of multiple publicly available pathway databases and 2) the set of PPIs per HIV-1 protein derived from the Jäger et al. 2012 data – ie for all 18 HIV-1 proteins we used all genes found to be interacting partners as a single pathway, resulting in 18 pathways. This set of pathway definitions was specific to MM5. We ran these analyses only using the main 'HIV-Acquisition' phenotype since it appeared to produce the strongest gene-level association results.

For the first set of analyses using MSigDB_C2, we display the top 5 pathways for both MM5 and P2 in Table 3.2. Encouragingly, we find at least one pathway related to HIV ('Reactome HIV Infection') among the top results from either MM5 or P2. In terms of genome-wide significance however, here defined as having a p-value $<1.06\times10^{-5}$

(.05/4,722 pathways in MSigDB_C2), no pathways appear to cross this threshold. On the level of marginal significance, we do see some pathways with p-values reaching an order of 10^{-4}. For the second set of analyses using the Jäger et al. 2012 pathways, we display the results for each HIV-1 protein in Table 3.3. Unfortunately, no single HIV-1 protein appears to produce a significant result (p-value $<2.7\times10^{-3}$, .05/18), with our strongest result only reaching a 10^{-2} order of magnitude (REV, p-value=.0114). However, here we took a particularly broad approach by including all genes found to interact with a given HIV-1 protein; it is possible that there are more logical subsets of genes, such as gene complexes, to analyze per HIV-1 protein that might produce stronger signals.

3.3.5 Selecting and supplementing variants for follow-up genotyping

In the second phase of our study, we will attempt to further increase our power to detect associations by expanding our analysis to the full MACS cohort. Specifically, we will conduct follow-up genotyping using Illumina custom genotyping arrays, a platform which should allow us to target up to 30,000 variants. In collaboration with other HINT researchers using the MACS dataset, we will also run the Illumina Multi-Ethnic Genotyping Array (MEGA) on the full MACS cohort, a genome-wide genotyping array that is enriched for ancestry-specific markers. Therefore we will have both a targeted, enriched set of variants to further our HIV-Acquisition and AIDS-Progression analyses, as well as a more general set of genome-wide variants to provide further background information.

To select SNPs for follow-up genotyping, we first collected the non-synonymous variants that we analyzed across MM5 and P2 (n=9,173 SNPs). Since this group of variants was below the total capacity of our genotyping array, we supplemented our current target list in multiple ways. First, we included any non-synonymous SNPs that were not used in our original analyses (n=1,207 SNPs); these were predominantly variants discovered in individuals of ancestries other than non-White Hispanic, since our analysis only focused on non-White Hispanics. Second, for a small number of top genes in both arrays, we included any variant that was discovered in our dataset (eg upstream, downstream, intronic, and any exonic SNP; n=9,637 SNPs). We determined 'top genes' by sorting our gene-level results from the main HIV-Acquisition analysis by SKAT-O p-values and then computing q-values for each gene. We then manually chose a q-value threshold that separated out a space-efficient number of genes and variants. Third, we used the ExAC database[131] to find additional non-synonymous variants to include for a handful of top genes once again. ExAC data was downloaded and variants were mapped to the genes on each array, and once again we used our q-value thresholding approach to determine top genes and variants (n=8,023 SNPs). We chose to emphasize additional non-synonymous variants since our HINT collaborators are more prepared to immediately interrogate interesting non-synonymous SNPs than any other type of variant. Lastly, we used the published GTEX data[33] to include putative eQTL SNPs for every gene that was analyzed. For each gene, a GTEX SNP was included if it contained an eQTL association p-value $<10x^{-4}$ from the whole-blood dataset. Additionally, if a secondary SNP >20kb away also passed this p-value threshold (distance being used as a proxy

for LD), it was also included (n=985). For further details regarding these steps see Methods, and for variant counts per array see Supplementary Table 3.7.

In total, we developed a set of 29,021 variants to be included on the Illumina custom genotyping array. Leftover room was filled by including other HINT research groups' targets of interest. At the current time, the arrays are in creation and we anticipate genotyping to begin in late Fall 2017.

3.4 Discussion

Here we present results from the first stage of our candidate gene-exome sequencing project looking at HIV-Acquisition and AIDS-Progression. We sequenced a subset of individuals from the MACS cohort targeting the exomes of genes with prior evidence of being related to HIV. We produced a set of high-quality variants numbering 149,063 in total across two arrays, and conducted gene-level rare variant tests of association for both HIV-Acquisition and AIDS-Progression. Overall, we found an enrichment of marginally significant p-values as compared to a permuted background of test statistics when looking at seronegative vs. seropositive individuals, and used the gene ranking from these analyses to inform the creation of follow-up genotyping custom arrays. We are now waiting for the data from the second stage of our study, which involves running these custom arrays as well as the Illumina MEGA chip on the full MACS cohort. We anticipate a greater signal of association for our main HIV-Acquisition analyses and aim to discover other positive signals in the other

approaches we conducted here.

Interestingly, when this study was first designed it was viewed as a cost-effective means to evaluate a set of candidate genes related to HIV. However, since the beginning of this project the discussion surrounding how best to approach association studies has expanded. Various other sequencing approaches continue to become cheaper and more accessible, including whole-exome, whole-genome, and now even single-cell sequencing. There are also recent methodological advancements that suggest much can be accomplished with even extremely low-coverage data[77]. Even the field's current view of genetic architecture is meaningfully being challenged, with a recent high-profile suggestion that for some traits the vast majority of the genome could play a relevant role[21]. With both our technologies and our understanding of biology continuing to improve, what actually entails the most 'cost-effective' approach for association studies will change as well; the best way to evaluate these different possibilities will be to carry out the experiments themselves. The second phase of this project is likely to provide insight into how targeted genotyping fits in with the current discussions. Therefore we not only hope to gain biological insight with the final results from this study, but also experimental insight into whether this approach is a viable possibility for future association studies as well.

Lastly, we note an ongoing part of this project has been parallel discussions with HINT collaborators. HINT researchers have already begun preliminary experiments using the top hits from the main HIV-Acquisition analysis of the first phase. Excitingly, initial results suggest that one of our top loci may be functionally related to

HIV-infection; a rare-variant discovered in our cohort appears to disrupt the binding between this locus and its HIV-1 protein partner. Additionally, presence of this rare-variant in a mutant copy of the gene appears to decrease HIV-infection rates in transfected cells. More follow-up work is necessary to confirm these initial results, but they point to the possibilities this study holds by directly combining sequencing and computational analysis with functional molecular work. We anticipate identifying more worthwhile candidate genes through the data collected in the second phase of our study and by working closely with our collaborators.

Overall, we aim to produce a complete study that identifies multiple, novel factors involved with HIV-Acquisition, as well as provide the initial stages of molecular follow-up of these loci. We also aim to show whether the study design used here is a viable and worthwhile approach in general for human genomics researchers. We anticipate exciting results either way with the completion of the second phase follow-up genotyping, as well as producing a resource for others' use from the MACS cohort.

3.5 Methods

3.5.1 Cohort

Samples were selected from the Multicenter AIDS Cohort Study (MACS)[110], a multicenter, prospective study of the natural and treated history of HIV-1 infection in at risk men who have sex with men. A total of 7,087 men have been enrolled since 1984 across four principal research sites: Baltimore MD, Chicago IL, Los Angeles CA, Pittsburgh PA. Each subject attends a semiannual visit for collection of clinical, demographic, psychosocial, behavioral and laboratory data. Banked specimens have been collected at each visit.

3.5.2 Sample Selection

We selected 988 individuals as a subset from the MACS cohort as follows:

- Highly-exposed seronegatives (n=200): participants with >45 high risk sexual encounters (anal receptive intercourse) in 2.5 years prior to their visit 2 and remained seronegative after 1990

- Low-exposed seronegatives (n=205): participants who reported <20 high risk sexual encounters in the 2.5 years prior to their visit 2 and remained seronegative after 1990

- Low-exposed seroconverters (n=181): seronegative participants who became infected in 1990 or earlier and who reported <20 high risk sexual encounters in the 2.5 years prior to their visit 2

- Other seroconverters (n=402): seronegative participants who became infected and who are not categorized as low-exposed

- Rapid and very rapid progressors (n=94): seropositive participants whose time to AIDS was within 5-years and 3-years of seroconversion, respectively.

- Slow and very slow progressors (n=146): seropositive participants who did to develop AIDS while being free of anti-viral therapy for at least 12-years and 15-years of seroconversion, respectively.

AIDS was defined as an absolute CD4 lymphocyte count of <200 or development of an AIDS-defining opportunistic infection or malignancy. For ancestries of these 988 individuals see Supplementary Table 3.4. For final, post-QC individual counts used in each analysis for both MM5 and P2 see Supplementary Table 3.5.

3.5.3 Gene Selection For NimbleGen Arrays

We selected candidate genes for sequencing through a two-step process. First, HINT researchers collected results from multiple previous studies to identify genes with strong experimental evidence for being functionally related to HIV. These studies included a genome-wide mass-spec protein-protein interaction (PPI) study[104], two

RNAi screens[22, 117], an unpublished, in-house RNAi screen, and an unpublished, in-house cDNA overexpression gain-of-function analysis to complement this suite of RNAi information. Genes were then ranked based on the number of times they appeared as a result in any of these studies; the more times a gene appeared as a result, the greater its rank.

Second, to focus and validate this list, the final rankings were compared against external collections of HIV- and immune-related genes, including 3,000 immune signaling and response genes in the Innate Immune Database (IIDB)[118], and interferon-stimulating genes and retrovirus restriction genes that are known to have undergone strong positive selection in primates[222]. 137 innate immune genes (conservatively defined as being present in 3 or more functional genomic screens) with previously unknown roles for HIV pathogenesis were strongly enriched among these databases. Included among our list of candidate loci that overlapped with these external resources were multiple restriction factors that are already known to interfere with HIV replication (eg, BST–2, APOBEC3G, and TRIM5[223, 51, 209]). This therefore suggests the functional experiments used to guide our gene selection were informative, and that our candidate gene list may be enriched for previously unknown loci that are functionally related to HIV.

These two steps, plus additional manual curation afterwards, led to the creation of a final list including 1,693 genes.

3.5.4 Sample Sequencing

High quality genomic DNA was extracted from stored peripheral blood mononuclear cells, buffy coat or cryopreserved, non-transformed lymphocytes. Two NimbleGen SeqCap EZ liquid capture arrays were used for target enrichment and long-read sequencing was accomplished on one of three native Roche 454 GS FLX+ pyrosequencers using the Titanium library chemistries. Standard quality checks and assurances were completed per protocol.

3.5.5 Sequence Mapping & Variant Calling

(Note – the following details where heavily influenced by the protocols and recommendations from 1000Genomes[73] and the Broad Institutes Genome Analysis ToolKit (GATK)[150, 43, 224])

To map sequencing runs, sff files were first converted to fastq format using sff_extract from the seq_crumbs package. Fastq files were quality controlled (QC'ed) by dropping reads that were either too short (<30bp) or too long (mean length + 6*SD length) via prinseq-lite 0.20. Filtered fastq files were then mapped to the 1000 Genomes human reference file version 37 (human_g1k_v37, eg hg19) via BWA 0.7.4 MEM[134]. Resulting sam files were converted to bam format and then QC'ed by dropping reads with low quality scores (MAPQ <10), were unmapped, or were secondary alignments using Samtools 0.1.19[135]. Filtered bam files were then merged across regions & runs for each individual sample using Picard 1.91 (ie if more than one sequencing run was

conducted for an individual, it was at this stage that these separate runs were now merged). Finally, per sample bam files went through Picard 1.91's duplicate removal and GATK 2.5.2's Base Quality Score Recalibration (BQSR).

In preparation for variant calling, all per sample bam files were merged using Picard. This single, merged bam file was then split by chromosome and processed through GATK's ReduceReads. Following this, variants were then called using GATK's Unified Genotyper according to GATK's Best Practices with the following dbsnp file and commands: 'dbsnp_137.b37.vcf -stand_call_conf 50 -stand_emit_conf 10 -dcov 250 -glm BOTH -nct 6'. Unified Genotyper was used over GATK's Haplotype Caller due to the known issue of 454 reads containing homopolymers; the presence of these homopolymers are problematic for haplotype-calling-based methods. The resulting per chromosome vcf files were merged back together to produce a single, main vcf file. This main variant file was first QC'ed using GATK's Variant Quality Score Recalibrator (VQSR) based on GATK's Best Practices and the following commands: '-nt 4 -percentBad .01 -minNumBad 1000
-resource:hapmap,known=false,training=true,truth=true,prior=15.0
hapmap_3.3.b37.vcf
-resource:omni,known=false,training=true,truth=true,prior=12.0
1000G_omni2.5.b37.vcf
-resource:1000G,known=false,training=true,truth=false,prior=10.0
1000G_phase1.snps.high_confidence.b37.vcf
-resource:dbsnp,known=true,training=false,truth=false,prior=2.0
dbsnp_137.b37.excluding_sites_after_129.vcf -an QD -an MQRankSum -an ReadPos-

RankSum -an FS -an DP -an HaplotypeScore -tranche 100.0 -tranche 99.9 -tranche 99.0 -tranche 90.0 -mode SNP'. SNPs were then removed if they fell outside the 99.9% tranche, were not biallelic, or were not within 1kb of a Nimblegen Custom array, using vcftools v0.1.11[36] and in-house code. Additionally, SNPs were removed if they had a missingness rate >5% or a Hardy-Weinberg Equilibrium p-value $<1e^{-4}$.

Individuals were then QC'ed using this refined list of called variants. First, Principal Component Analysis (PCA) was conducted using EIGENSOFT 5.0.1's smartpca[173] and autosomal SNPs. PCA outliers were removed by manual inspection. Individuals that only contained European ancestry were then kept for further analysis. Additionally these remaining individuals were QC'ed based on Identity-By-Descent (IBD) measures and genotype missingness; individuals were dropped in the presence of cryptic relatedness (high, outlier z1 values from PLINK[177] v1.07's --genome) and >20% missingness.

3.5.6 SNP Annotation

SNPs were annotated using ANNOVAR[234] (downloaded 9/11/2013) and the following generic command, 'perl summarize_annovar.pl InputFile1.txt /Path/To/Databases/ --ver1000g 1000g2012apr --verdbsnp 129 --veresp 6500 --buildver hg19 remove'. This command corresponds to the following databases and versions being used while mapping to Hg19: 1000Genomes (April 2012 release), dbSNP (129), ESP (6500). All other default ANNOVAR values and weights were used.

3.5.7 Overlap with 1000Genomes and ExAC

1000 Genome[73] data was accessed by downloading the 3rd release vcfs from the 1000 Genome's UK portal (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/). European-ancestry allele frequency information was extracted from the 1000 Genomes data using the 'Eur_AF' field per SNP ('Eur_AF' field description from the 1000 Genomes vcf: Allele frequency in the EUR populations calculated from AC and AN, in the range (0,1)). SNPs that did not have information in this field and SNPs that were non-biallelic were dropped. ExAC[131] data was accessed by downloading the main vcf of the 3.0 release from the ExAC website (ftp://ftp.broadinstitute.org/pub/ExAC_release/release0.3/). European-ancestry frequency information was calculated from the ExAC data by first extracting the AC_NFE and AN_NFE fields ('AC_NFE' and 'AN_NFE' field descriptions from ExAC vcf: Non-Finnish European Allele Counts & Non-Finnish European Chromosome Count). ExAC per SNP allele frequency was then determined by AC_NFE / AN_NFE. SNPs that did have information in either of these fields were dropped.

The SNP overlap between these datasets and our set of called variants were determined via in-house scripts. Allele frequencies per SNP were then plotted against one another using R.

3.5.8 SKAT-O Analysis & Permutations

SKAT-O[242, 130] analyses were conducted using the software suite EPACTS v3.2.4 (http://genome.sph.umich.edu/wiki/EPACTS) via the following generic command: 'epacts --vcf InputFile1.vcf --groupf InputGeneVariantList.txt --ped InputFile1.ped --pheno Phenotype* --test skat --skat-o --skat-adjust --beta 1,25 --max-maf 1.0 --cov RMID --cov PC1 --cov PC2 --cov PC3 --cov PC4 --cov PC5'. Where 'Phenotype*' refers to any of the four below phenotype categories, 'InputGeneVariantList.txt' is a list of which variants to include per gene for the SKAT-O gene-based test, and covariates 'PC1-5' and 'RMID' refer to the top 5 PCs as determined during the earlier individual QC section and the ID# specific to the 454 MID adapter used in the sequencing process for that individual (ranges from 1-12). Phenotype groups for case/control analyses were constructed as follows: Seronegative vs. Seropositive ('HIV-Acquisition'), Highly-Exposed Seronegative vs. Seropositive ('HIV-Acquisition HE'), Very Rapid + Rapid Progressors vs. Very Slow + Slow Progressors ('AIDS-Progression'), and Very Rapid Progressors vs. Very Slow Progressors ('AIDS-Progression Extreme'). SKAT-O tests were conducted using all the non-synonymous variants called within the QC'ed white non-Hispanic subset for a given gene.

Permutations were conducted by randomly shuffling the phenotype designations across individuals. This was done 1,000 times and SKAT-O analyses for each of the four case-control groups described above were conducted on each individual permutated dataset. QQ-plots were created by comparing the gene-rankings of the

original SKAT-O results to the gene-rankings of the permuted results, eg the top gene's p-value in the original results was compared against the top gene's p-value in all 1,000 permuted datasets.

3.5.9 Pathway Analysis

Pathway analysis was conducted using SKAT-O in a manner similar to that described above. The main change was how we defined the loci being tested. Instead of defining a 'locus' for tests as all the non-synonymous variants in a given gene, a 'locus' was defined as all the non-synonymous variants across the multiple genes a given pathway includes. This refers specifically to the groups defined in the --groupf file InputGeneVariantList.txt. To define pathways for these SKAT-O analyses, we used two approaches. First, the Broad Institute's MSigDB C2 collection was used to define a generic collection of known and curated pathways (downloaded from http://software.broadinstitute.org/gsea/msigdb/collections.jsp). Second, previous HIV studies conducted by collaborators were used to determine different candidate sets of genes and pathways. These were based on previous siRNA screens[22, 117], proteomics work (Jäger et al. 2012), and other internal research efforts.

3.5.10 SNP Selection for Follow-up Genotyping

Variants were chosen for follow-up genotyping based on the following five categories. For specific details regarding number of genes and variants included per category

based on the below criteria, see Supplementary Table 3.7:

- *SKAT-O HIV Non-Synonymous Variants*: All non-synonymous variants used in the original set of SKAT-O tests were included for follow-up genotyping

- *Remaining HIV Non-Synonymous Variants*: Non-synonymous variants that were called but not included in the final SKAT-O analyses; these were generally non-synonymous variants that were identified in individuals outside of the white non-Hispanic ancestry.

- *Top-Gene HIV All Variants*: For a small set of top genes in the first array, all variant-types called in the HIV dataset were included. Specifically, genes were first ranked by SKAT-O p-value and then q-values were determined using the R package 'qvalue'. Genes whose q-values were $\leq$ a specified threshold were then included. This lead to all variant types discovered (eg intronic, exonic, 5'/3' UTR, upstream/downstream) in these genes being included. For MM5 a q-value cutoff of .46 was used, and for P2 a q-value cutoff of .52 was used.

- *Top-Gene ExAC Non-Synonymous Variants*: Similarly, for a small set of top genes in both arrays, the ExAC dataset[131] was used to identify additional non-synonymous variants to include per gene. To determine which ExAC SNPs were non-synonymous for a given gene, variants were annotated using ANNOVAR as described earlier. For MM5 a q-value cutoff of .4 was used, and for P2 a q-value cutoff of .5 was used. ExAC release0.3.1 was used (downloaded from ftp://ftp.broadinstitute.org/pub/ExAC_release/release0.3.1/).

- *GTEx Whole-Blood eQTL Variants*: To further supplement our follow-up genotyping list, we used the GTEX data[33] to include putative whole-blood eQTL variants corresponding to the genes from our SKAT-O analyses. We used two steps to determine whether to include up to two variants per gene as putative eQTLs. First, we identified whether the SNP with the strongest signal had a p-value a p-value $<1e^{-4}$. Second, we then identified whether there was a second SNP that was at least outside a 20kb window of this first SNP that also had a p-value $<1e^{-4}$. This second step was an attempt to identify a secondary signal outside of a possible local LD block. For a given gene then, it could have anywhere between 0, 1, or 2 additional SNP included for follow-up genoptyping depending on whether either of these two criteria were met. To conduct this setup, GTEx V6 data was used (downloaded from http://www.gtexportal.org/home/datasets).

3.6 Acknowledgments

We thank Bryan Howie, Audrey Fu, and other members of the Stephens Lab for helpful input at various points during the implementation of this project. We also John Zekos for support, as well as John Novembre, Anna DiRienzo, Xin He, and other members of the Cummings 4$^{\text{th}}$ Floor for periodic feedback with the first phase of this study. Additionally, we thank Aleksandar Zivkovic and other members of the Wolinsky Lab for continued support throughout the project. Finally, we thank Kevin Olivieri, Sumit Chanda, Nevan Krogan, and other members of HINT for their continued involvement and collaboration. This work was supported by National Institutes of Health Grant U01 AI035039 to SMW (subcontract to MS), and NIH Grants T32 GM007197 & F31 AI118375 to MCT.

3.7 Author Contributions

SP, MS, and SMW conceived the original study design. SP, EK, and KK conducted the sample preparation and DNA sequencing. MCT conducted the sequence-mapping and variant-calling. MCT and SP performed the analyses. SP, MS, and SMW supervised the project. MCT wrote the manuscript with input from SP, MS, and SMW.

3.8 Figures

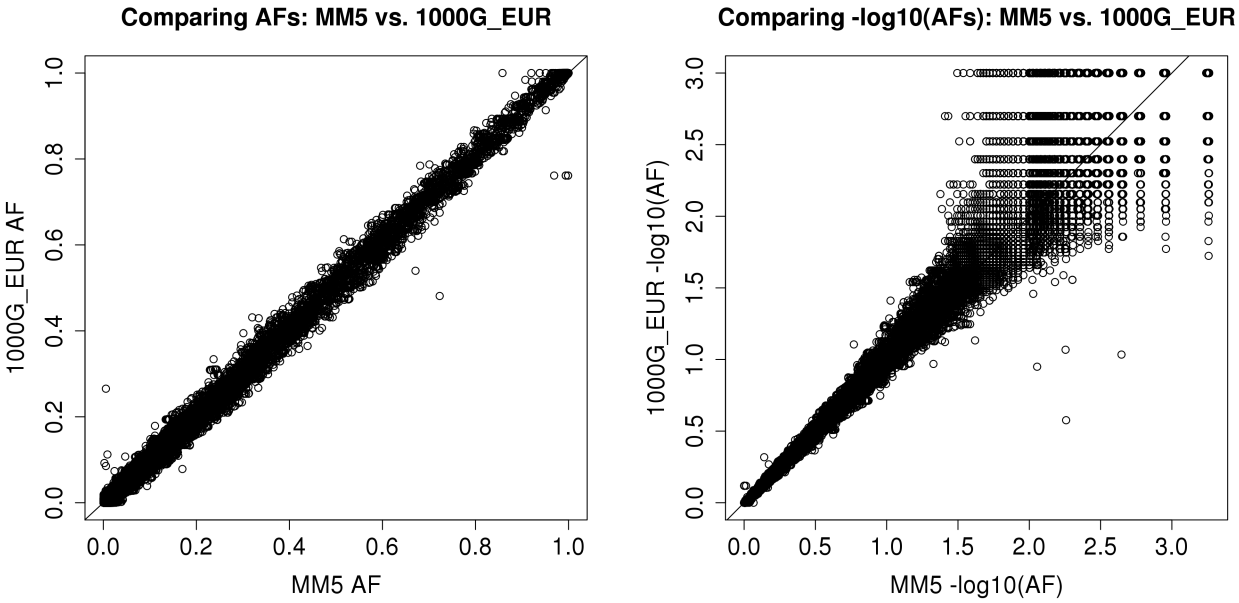

Figure 3.1: **MM5 Overlap with 1000G** – Shown are plots comparing the allele frequencies (AF) of overlapping variants between the MM5 post-QC data and 1000 Genomes (1000G) data from the European subset. The first plot shows original AFs being compared and the second plot shows -log10(AFs) being compared. 503 individuals were used from the 1000G dataset, and a total of 39,049 SNPs were found to overlap between the two datasets. See Methods for further details on how 1000G data was extracted.

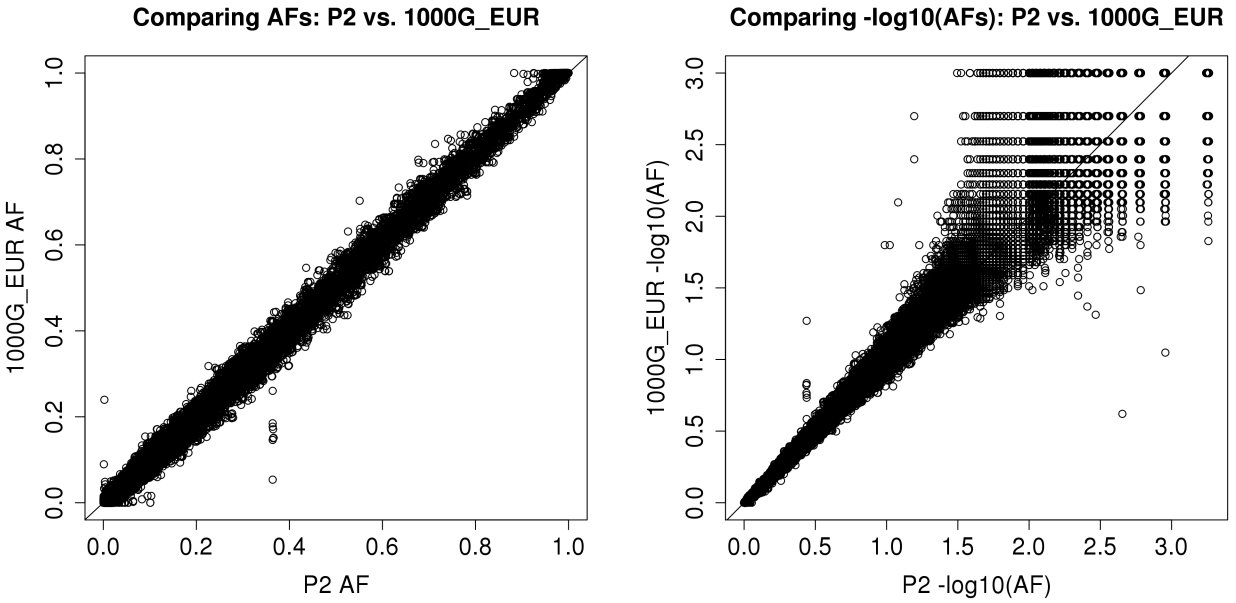

Figure 3.2: **P2 Overlap with 1000G** – Shown are plots comparing the AFs of overlapping variants between the P2 post-QC data and 1000 Genomes (1000G) data from the European subset. A total of 54,910 SNPs were found to overlap. See Figure 3.1 description for further details.

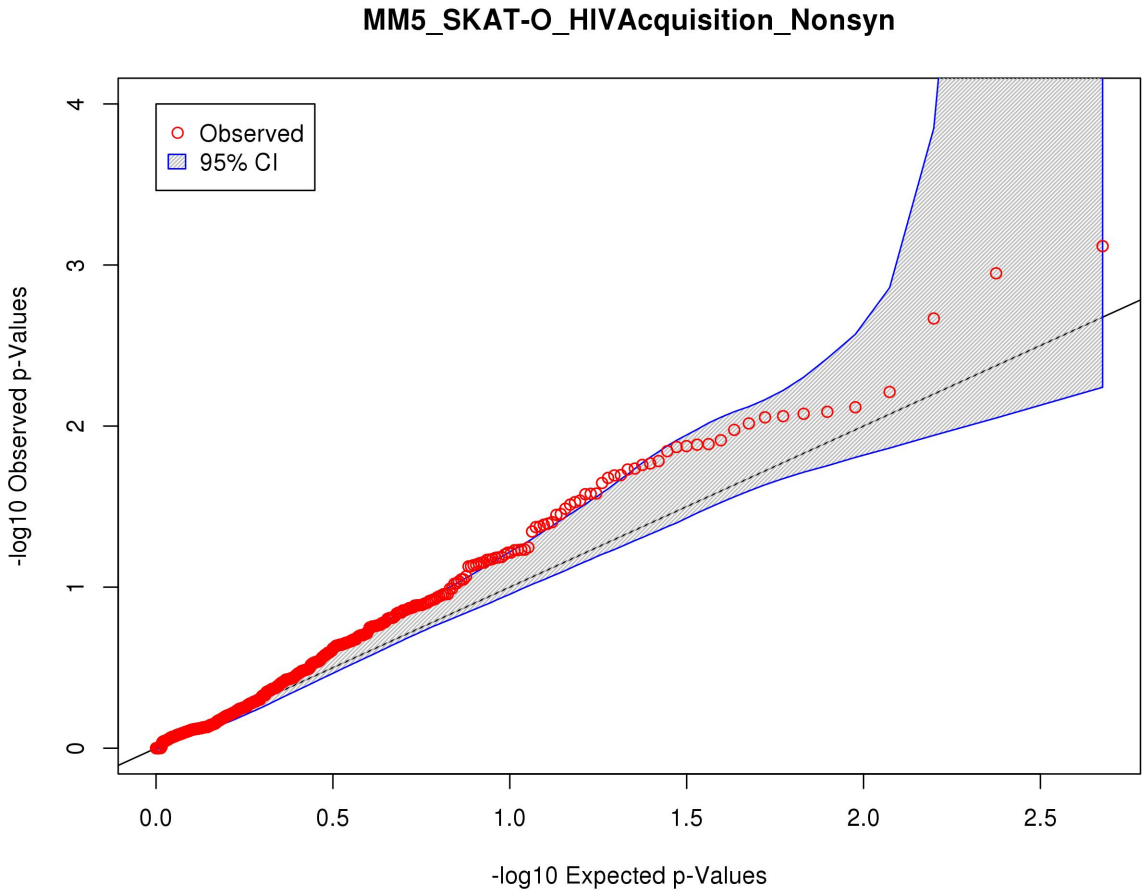

Figure 3.3: **MM5 SKAT-O QQPlot: HIV-Acquisition** – Shown is a Quantile-Quantile Plot (QQPlot) displaying the results for the SKAT-O analyses of the HIV-Acquisition phenotype from the MM5 array. SKAT-O analyses are conducted on a 'per- locus' level, therefore datapoints represent variants grouped together and not individual variants. The x-axis is the -log10 of the expected SKAT-O p-values and the y-axis is the -log10 of the observed SKAT-O p-values. Observed, true results are shown in red, whereas a 95% confidence interval derived from 1,000 permutations is shown in gray. Only non-synonymous variants were used per-gene for this analysis. The HIV-Acquisition analysis was conducted using seronegative individuals vs. seropositive individuals.

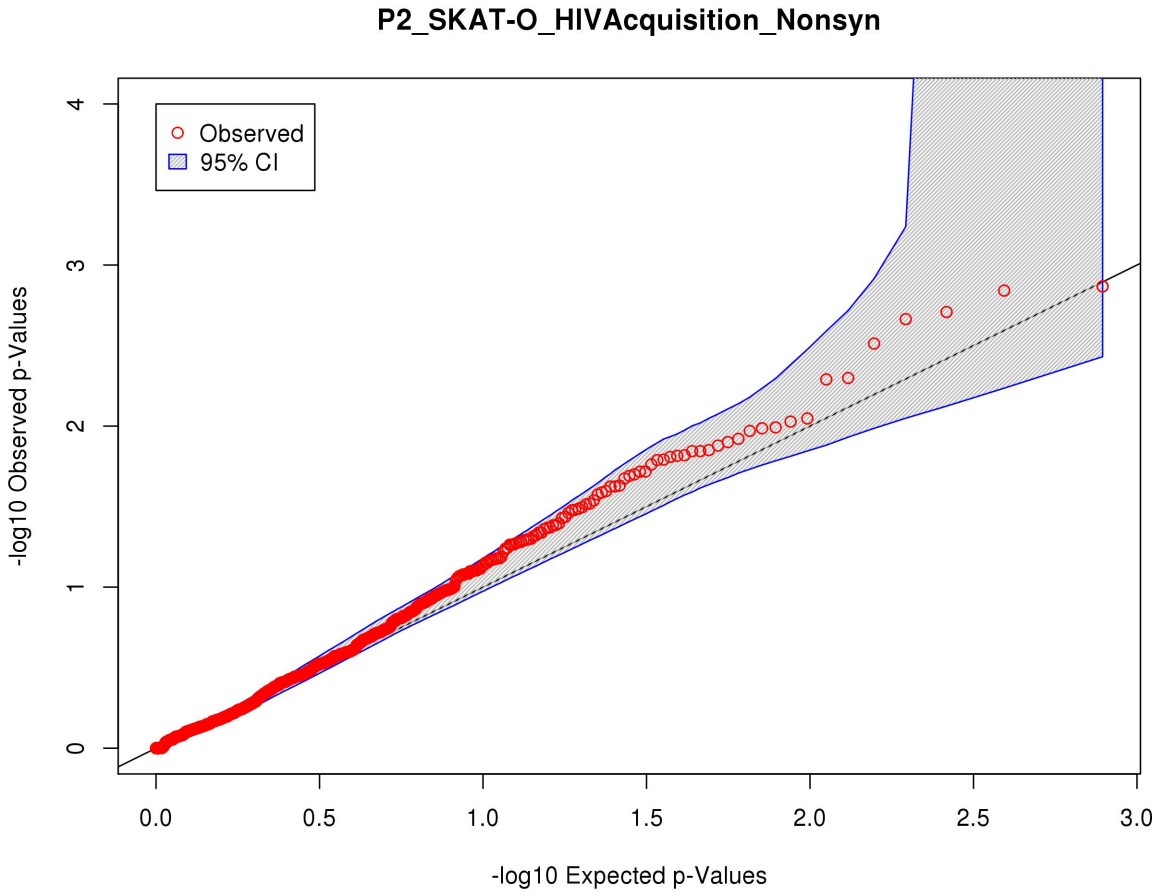

Figure 3.4: **P2 SKAT-O QQPlot: HIV-Acquisition** – Shown is the QQPlot for the SKAT-O analyses of the HIV-Acquisition phenotype from the P2 array. See Figure 3.3 for the remaining details.

3.9 Tables

Gene	p-Value	q-Value	# Vars	Rho
MM5				
GEMIN2	7.64×10^{-4}	0.209	4	0.09
VTI1A	1.13×10^{-3}	0.209	4	0.01
VDAC3	2.15×10^{-3}	0.266	5	0
SDHD	6.14×10^{-3}	0.311	1	NA
NUDC	7.64×10^{-3}	0.311	4	0.04
P2				
XAF1	1.36×10^{-3}	0.395	7	1
PRDM1	1.44×10^{-3}	0.395	12	1
PDXK	1.96×10^{-3}	0.395	3	0.09
MAPK14	2.17×10^{-3}	0.395	3	1
NLRX1	3.08×10^{-3}	0.448	19	0.04

Table 3.1: **Top SKAT-O HIV-Acquisition Results: MM5 & P2** – Shown is a table of top SKAT-O results for both the MM5 and P2 arrays. Specifically, the top 5 loci are shown from the SKAT-O analysis of HIV-Acquisition (seronegatives vs. seropositives). 775 individuals were included in the MM5 analysis and 747 individuals were included in the P2 analysis. The first column shows the gene ID, the second column shows the SKAT-O p-value, the third column shows the associated q-value, the fourth column shows the number of non-synonymous variants used for the gene, and the last column shows the rho value from the SKAT-O analysis. The SKAT-O rho value represents the mixture proportion between the 'variance' and 'burden' tests, with 0 being 100% the 'variance' test and 1 being 100% the 'burden' test. q-values were determined from the full distribution of SKAT-O p-values. For full details on the SKAT-O analyses, see Methods.

Pathway	p-Value	# Genes Pathway	# Genes Dataset	# Vars Dataset	Rho
MM5					
Zheng Bound By FOXP3	5.63×10^{-4}	491	14	104	0.09
Kohoutek CCNT1 Targets	1.26×10^{-3}	50	2	14	1
Kayo Aging Muscle Up	1.48×10^{-3}	224	6	113	0
REACTOME HIV Infection	2.04×10^{-3}	207	18	136	0
Kaab Heart Atrium VS Ventricle Dn	2.39×10^{-3}	261	13	52	0.09
P2					
Dazard UV Response Cluster_G3	1.84×10^{-4}	15	2	15	1
SU Pancreas	2.00×10^{-4}	54	2	26	0.25
Stambolsky Response To Vitamin_D3 Up	2.72×10^{-4}	84	9	49	0.09
Williams ESR1 Targets Up	2.87×10^{-4}	26	4	19	0.25
Shin B_Cell Lymphoma Cluster_2	1.80×10^{-3}	30	4	27	0.25

Table 3.2: **Pathway Analysis: MSigDB_C2** – Pathway analysis of both MM5 and P2 genes using SKAT-O and the Broad Institute's MSigDB C2 collection. The MSigDB_C2 collection represents a manually curated collection of publicly available pathway databases and resources, containing a total of 4,722 pathways. SKAT-O was used by specifying a single 'locus' as all the non-synonymous variants from all the genes present for a given pathway from MSigDB_C2. The first column displays the name of the pathway from the MSigDB_C2 collection. The second column displays the p-value from the SKAT-O analysis. The third column displays the total number of genes associated with the given pathway. The forth column displays the number of genes from our dataset that mapped to the given pathway. The fifth column display the number of non-synonymous variants used for the analysis across the given number of genes included. And the sixth column displays the rho value from the SKAT-O analysis; for a description of the rho value see Table 3.1 description and Methods.

HIV-1 Protein	p-Value	# Genes preQC	# Genes postQC	# Vars	Rho
REV	0.0114	14	14	74	0.0061
VPR	0.0667	40	37	221	0.04
NEF	0.137	8	8	41	1
VPU	0.145	55	50	515	0
TAT	0.175	23	19	192	0
VIF	0.273	22	18	104	1
GAG	0.339	5	5	26	0.8
PR	0.413	44	42	627	0
POL	0.525	31	28	222	1
GP120	0.536	43	41	253	0.25
RT	0.538	2	2	28	1
GP160	0.612	41	38	264	0
IN	0.656	19	16	205	0
NC	0.710	11	9	61	1
GP41	0.760	36	28	114	0
MA	0.916	18	17	159	0

Table 3.3: **Pathway Analysis: Jäger PPIs** – Pathway analysis of MM5 using SKAT-O and the Jäger et al. 2012 PPI data. As described in text, Jäger et al. 2012 produced a list of genes found to interact with each HIV-1 protein using a genome-wide immunoprecipitation and mass-spectrometry approach. These results were then used to construct pathways by collecting every gene on MM5 that was found to associate with each of the HIV-1 proteins, resulting in 16 different pathways corresponding to 16 of the 18 HIV-1 proteins. Analysis was conducted on these pathways using SKAT-O in the same manner as described in the Table 3.2 description. The first column here shows the HIV-1 protein whose 'pathway' is being tested. The second column shows the resulting SKAT-O p-value from the analysis. The third column shows the original number of genes on the MM5 array that were associated with the specified HIV-1 protein, and the fourth column shows the number of genes that remained post-QC. The fifth column shows the number of non-synonymous variants that were used from the genes included for analysis, and the sixth column shows the SKAT-O rho value, which is detailed in the Table 3.1 description.

3.10 Supplementary Figures

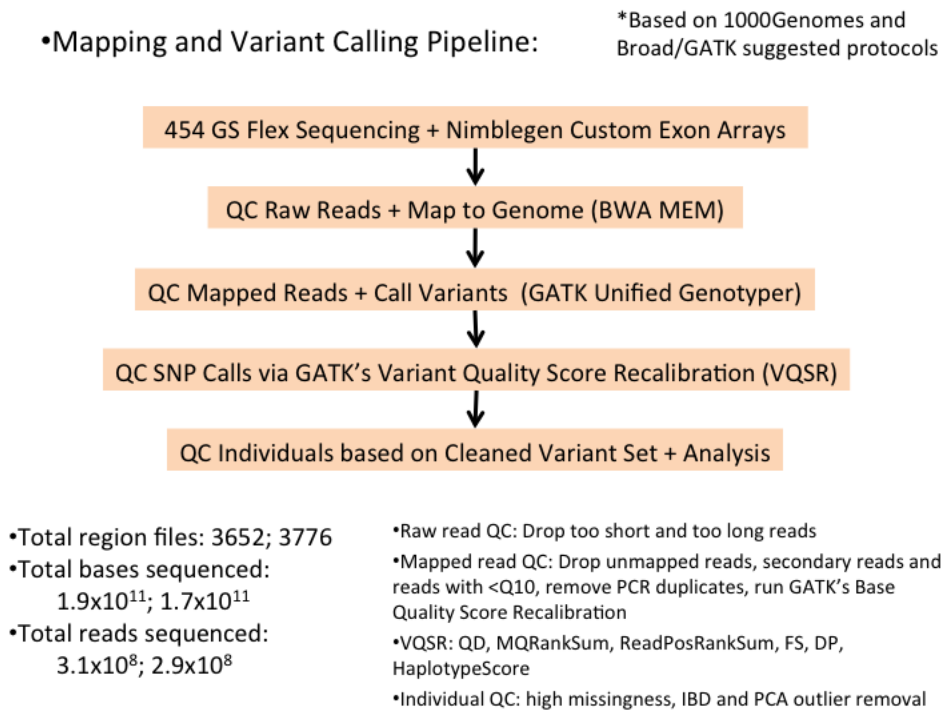

Figure 3.5: **Mapping & Variant Calling Pipeline** – Shown is a summary of the mapping and variant calling pipeline used to process the MACS subset targeted gene-exome sequencing data. In brief, mapping was conducted using the BWA MEM algorithm and variant-calling was conducted using GATK's Unified Genotyper. QC procedures were included at all stages of the pipeline, including pre-mapping, post-mapping, post-variant calling, as well as on the individual level. Some details can be found in the bottom of the figure; for full details of QC and all other parts of the pipeline see Methods. Additionally, as noted at the top, the pipeline used here is based on 1000Genomes and GATK Best Practices protocols. Summary information regarding total number of files processed, bases sequenced, and reads sequenced, for MM5 and P2 respectively, are also included on the bottom.

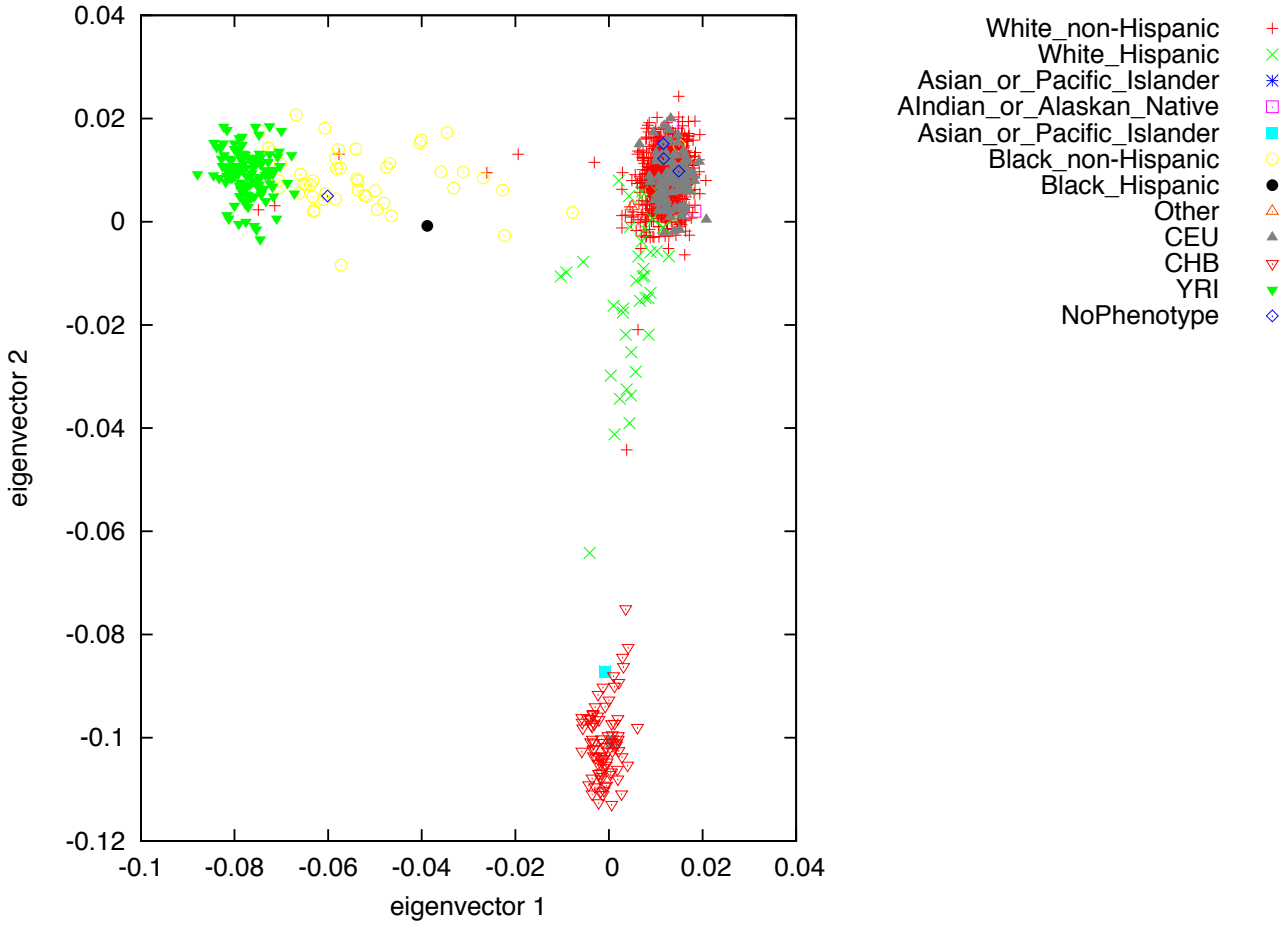

Figure 3.6: **MACS Subset PCA Plot** – Shown is a Principal Components Analysis (PCA) plot of the first two PCs from the MACS subset as well as individuals from the HapMap3 CEU, CHB, and YRI cohorts. The x-axis is the 1st PC and the y-axis is the 2nd PC. 3,902 overlapping autosomal SNPs between the MACS subset and HapMap3 individuals were used to perform the analysis. The plot displayed here is based on the MM5 data; similar results were found from using P2 data. Outliers from the main White non-Hispanic cluster were removed as part of QC. The top 5 PCs were used as covariates in the SKAT-O analyses. See Methods for individual counts per ancestry and figure legend for plot point details.

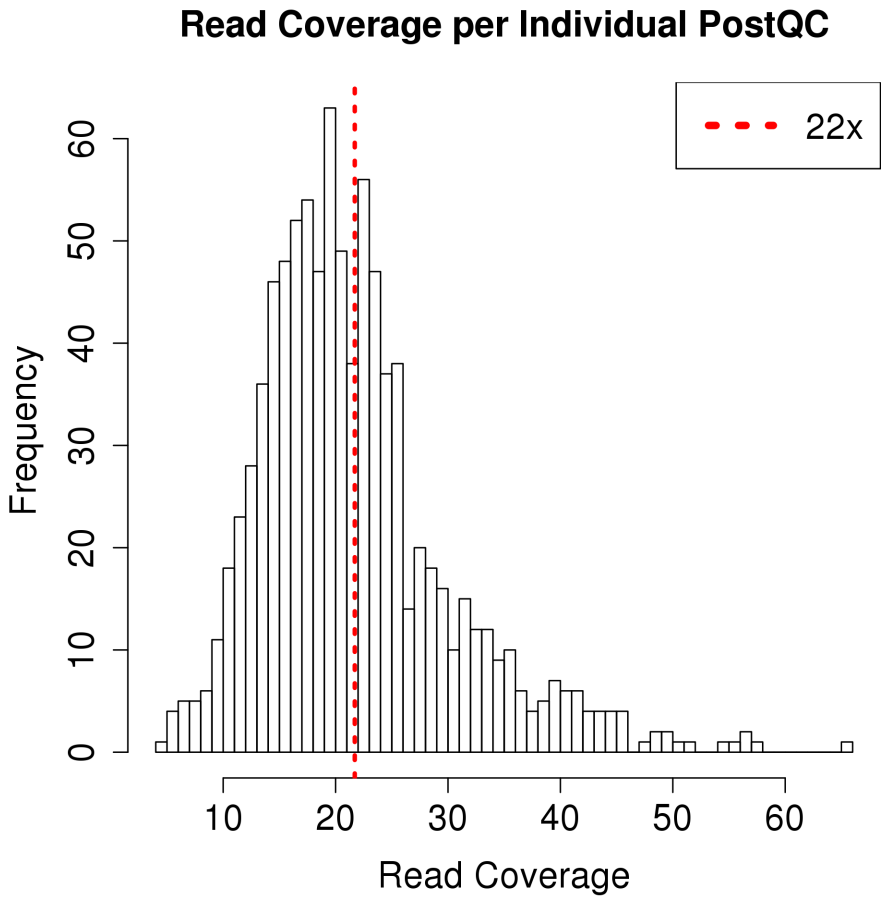

Figure 3.7: **MM5 PostQC perIndv Coverage Histogram** – Shown is a histogram of the mean fold read coverage per individual post-QC. Data presented here is from the MM5 array. All individuals post-QC, pre-analysis are included. The red dashed line is placed at the average fold coverage across all individuals, with the value shown in the top-right legend.

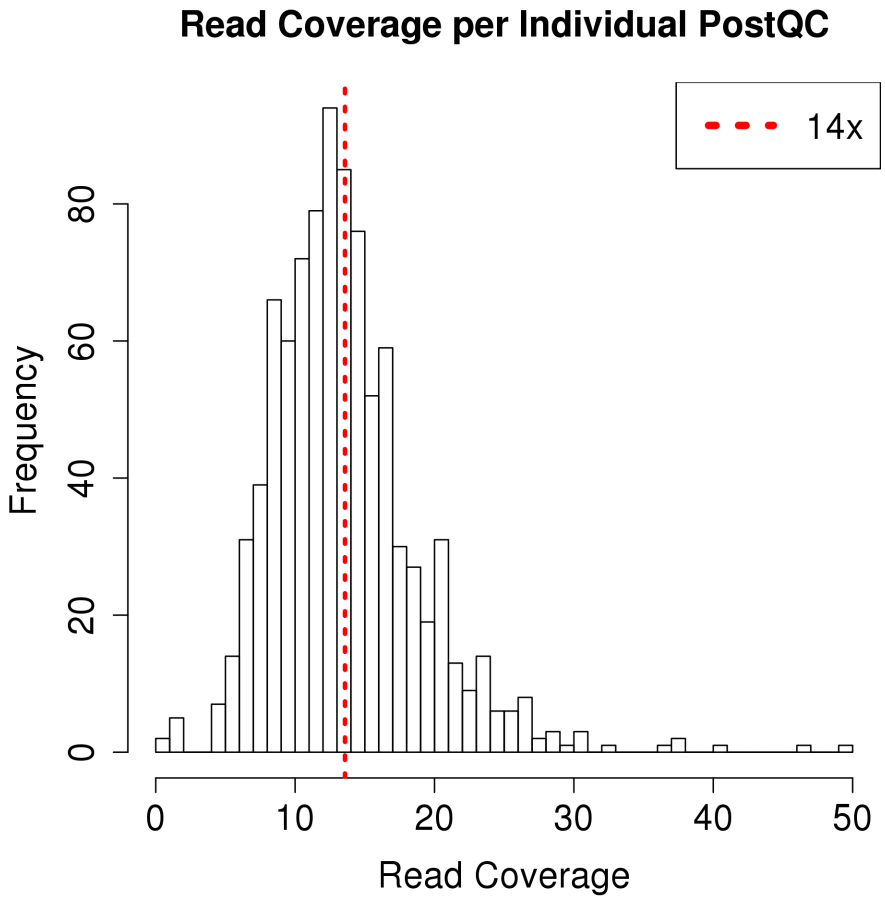

Figure 3.8: **P2 PostQC perIndv Coverage Histogram** – Shown is a histogram of the mean fold read coverage per individual post-QC. Data presented here is from the P2 array. All individuals post-QC, pre-analysis are included. The red dashed line is placed at the average fold coverage across all individuals, with the value shown in the top-right legend.

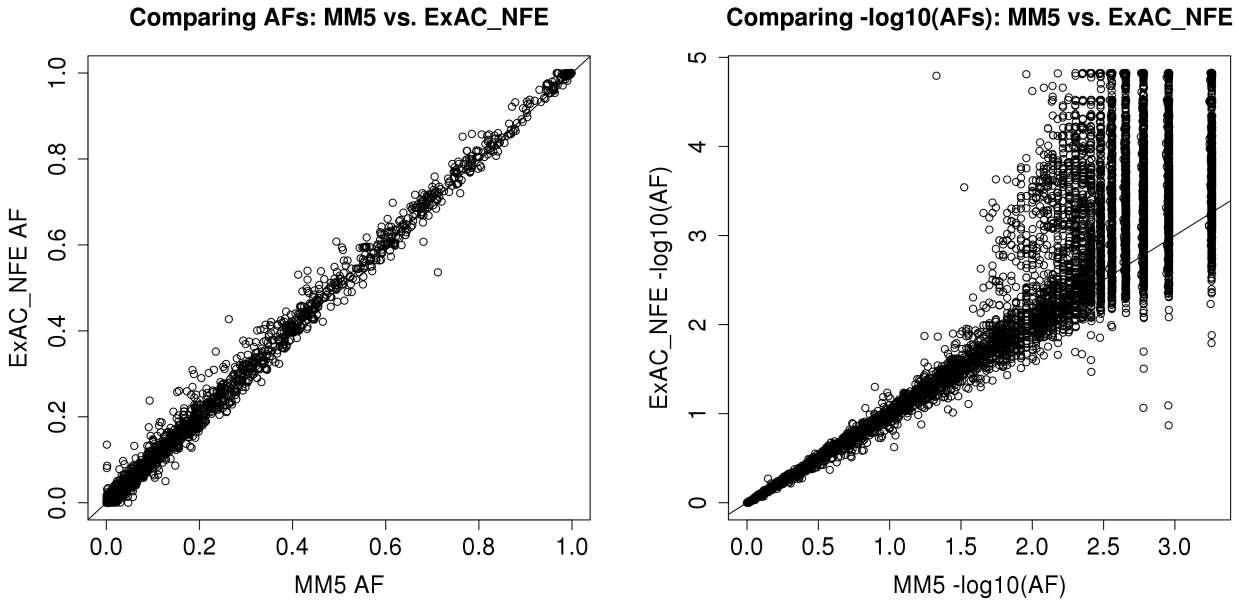

Figure 3.9: **MM5 Overlap with ExAC** – Shown are plots comparing the AFs of overlapping variants between the MM5 post-QC data and Exome Aggregation Consortium (ExAC) data from the non-Finnish European subset. An average of ∼30,000 individuals were used from the ExAC dataset, and a total of 11,339 SNPs were found to overlap. See Figure 3.1 description for further details and Methods for how ExAC data was extracted.

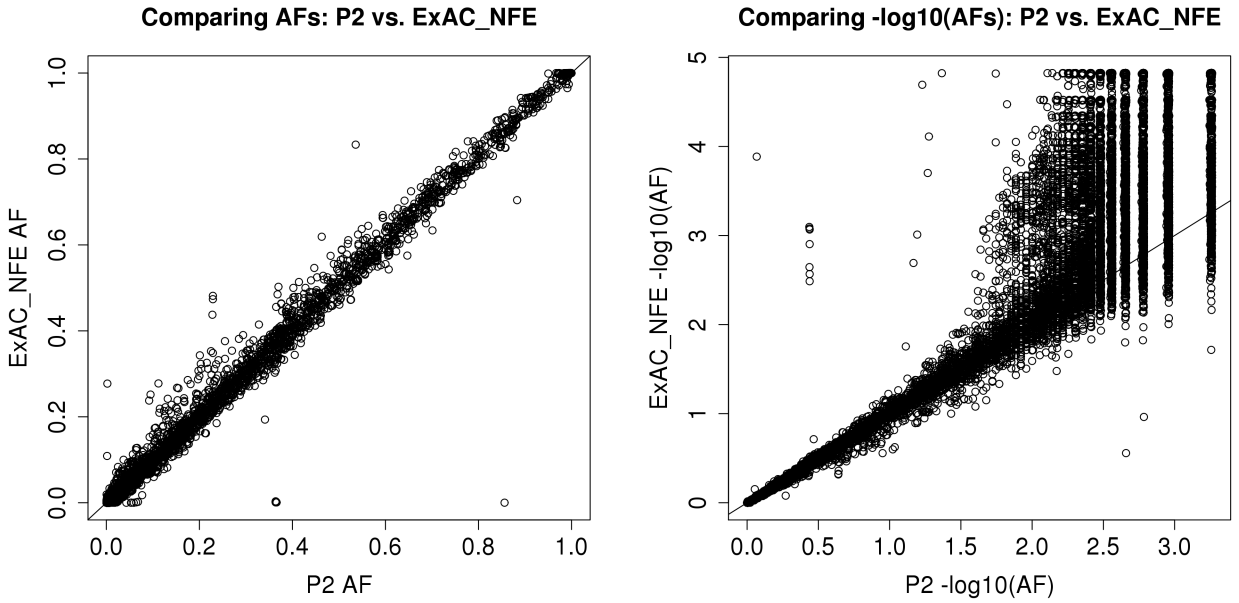

Figure 3.10: **P2 Overlap with ExAC** – Shown are plots comparing the AFs of overlapping variants between the P2 post-QC data and Exome Aggregation Consortium (ExAC) data from the non-Finnish European subset. A total of 15,774 SNPs were found to overlap. See Figure 3.1 and Supplementary Figure 3.9 descriptions for further details.

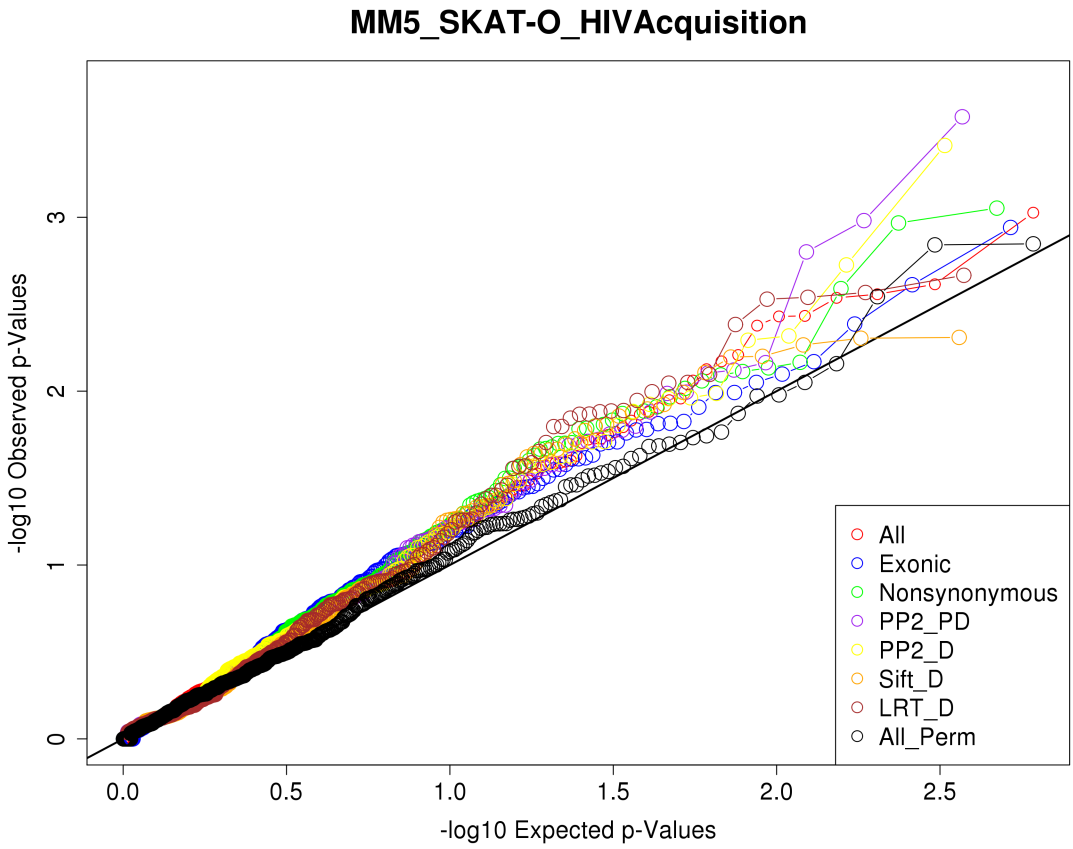

Figure 3.11: **MM5 Preliminary SKAT-O Analyses: HIV-Acquisition** – Shown is a QQPlot for preliminary analyses of SKAT-O exploring different design strategies for collecting variants per gene. SKAT-O requires users to define which variants to include per gene; therefore to identify the best approach different grouping possibilities were examined. Design strategies explored include: using all variants, using only exonic variants, using only non-synonymous variants, using only variants defined as 'possibly damaging' and 'probably damaging' by Polyphen2, using only variants defined as 'probably damaging' by Polyphen2, using only variants defined as 'damaging' by Sift, and using only variants defined as 'damaging' by LRT. The phenotype used here is HIV-Acquisition (seronegatives vs. seropositives). SKAT-O results from a single permutation are also shown as a comparison ('All_Perm'). The x-axis is the -log10 expected SKAT-O p-values and the y-axis is the -log10 observed SKAT-O p-values. See legend for color coding details.

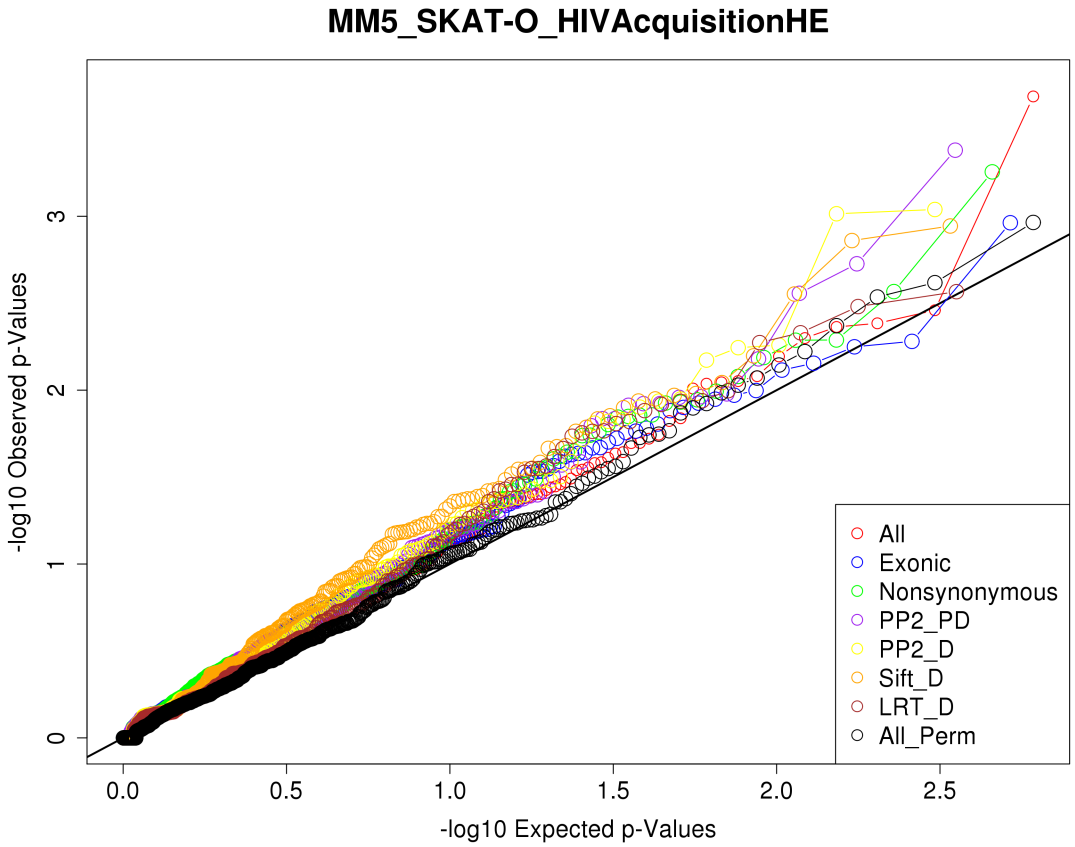

Figure 3.12: **MM5 Preliminary SKAT-O Analyses: HIV-Acquisition HE** – Shown is a QQPlot for preliminary analyses of SKAT-O exploring different design strategies for collecting variants per gene. The phenotype used here is HIV-Acquisition HE (highly-exposed seronegatives vs. seropositives). See Figure 3.11 for remaining details.

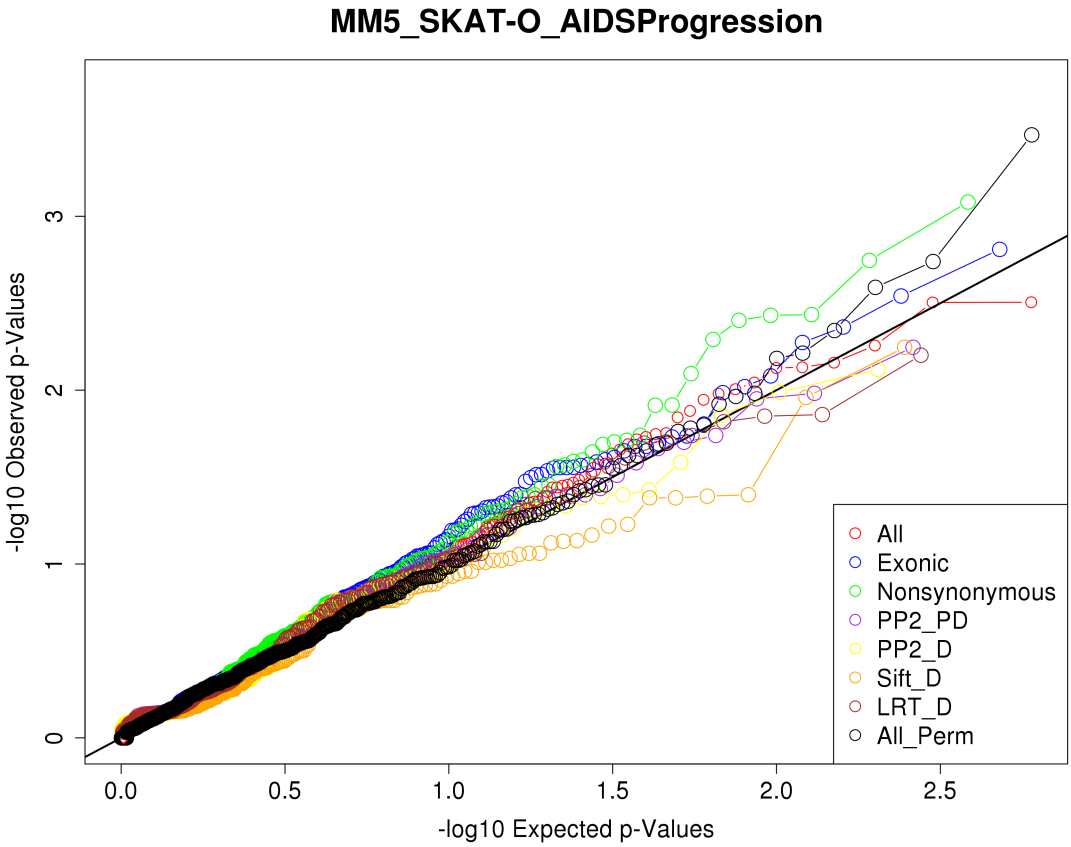

Figure 3.13: **MM5 Preliminary SKAT-O Analyses: AIDS-Progression** – Shown is a QQPlot for preliminary analyses of SKAT-O exploring different design strategies for collecting variants per gene. The phenotype used here is AIDS-Progression (very rapid + rapid progressors vs. very slow + slow progressors). See Figure 3.11 for remaining details.

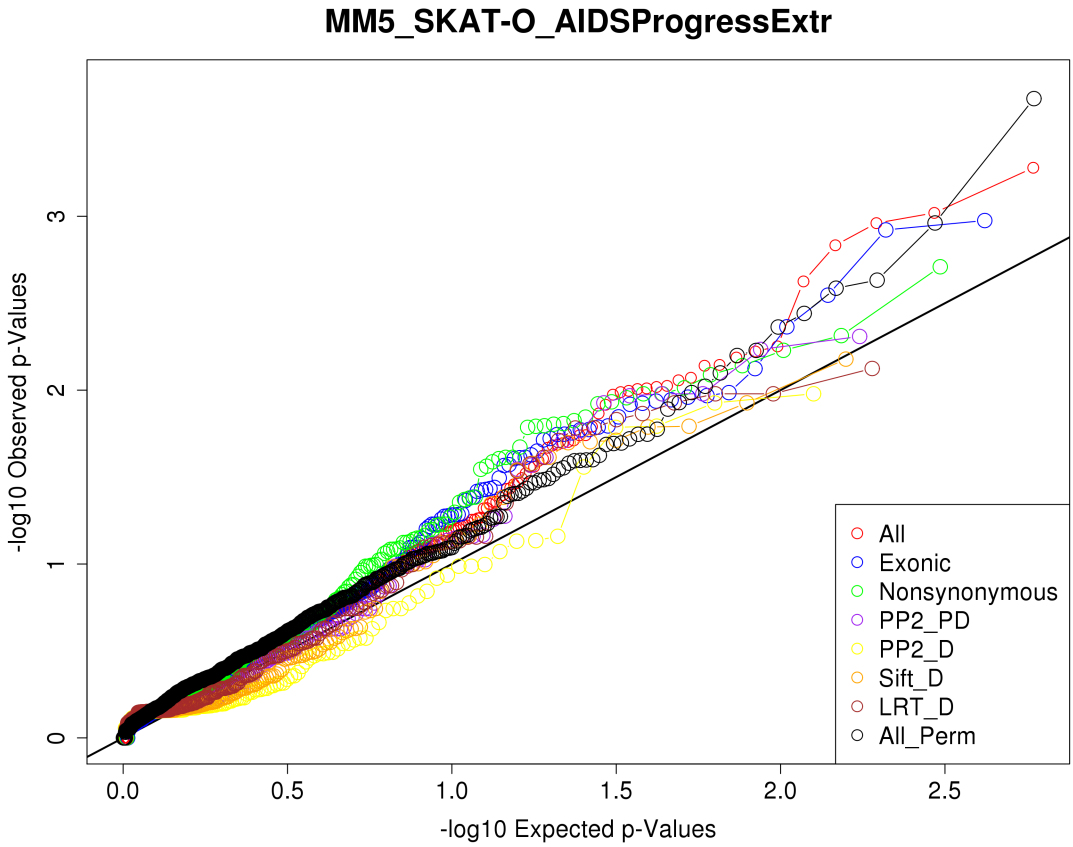

Figure 3.14: **MM5 Preliminary SKAT-O Analyses: AIDS-Progression Extr** – Shown is a QQPlot for preliminary analyses of SKAT-O exploring different design strategies for collecting variants per gene. The phenotype used here is AIDS-Progression Extr (very rapid progressors vs. very slow progressors). See Figure 3.11 for remaining details.

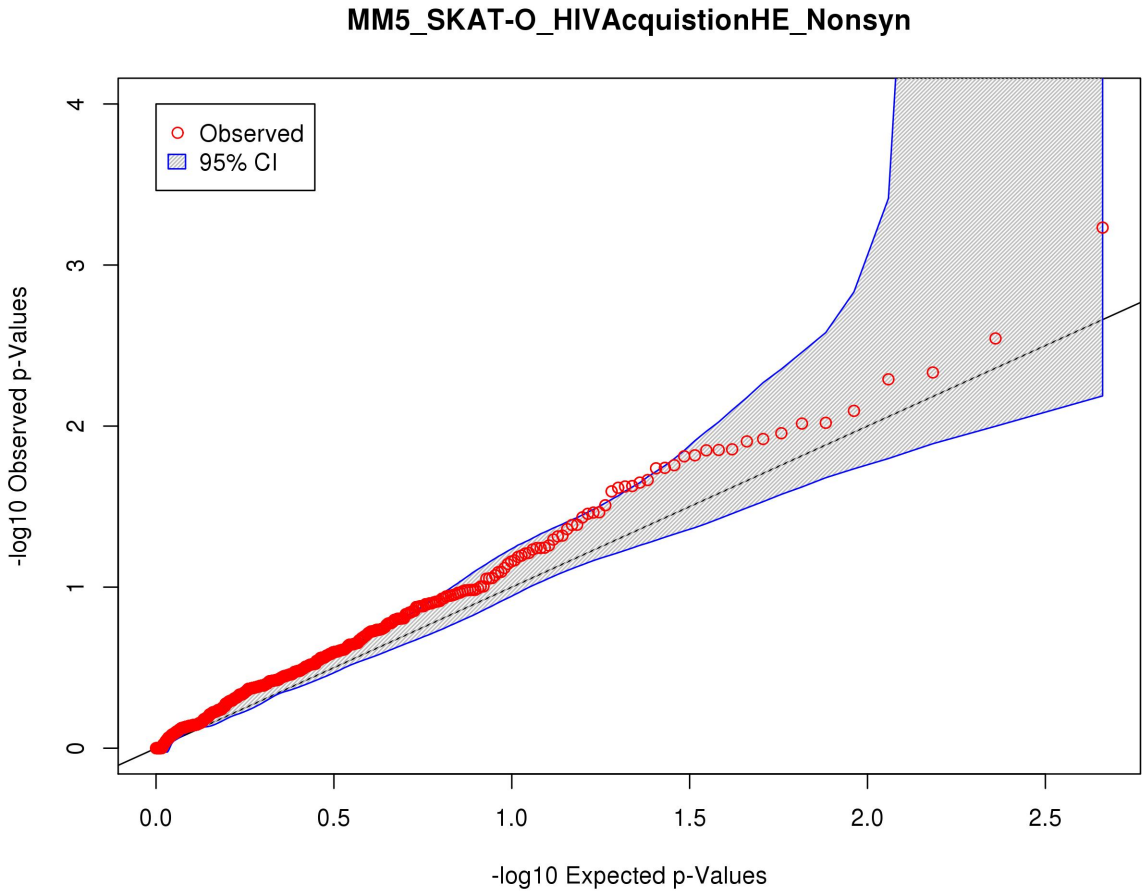

Figure 3.15: **MM5 SKAT-O QQPlot: HIV-Acquisition HE** – Shown is the QQPlot for the SKAT-O analyses of the HIV-Acquisition HE phenotype from the MM5 array. Here highly-exposed ('HE') seronegative individuals are used as the controls vs. the seropositive cases. See Figure 3.3 for remaining details.

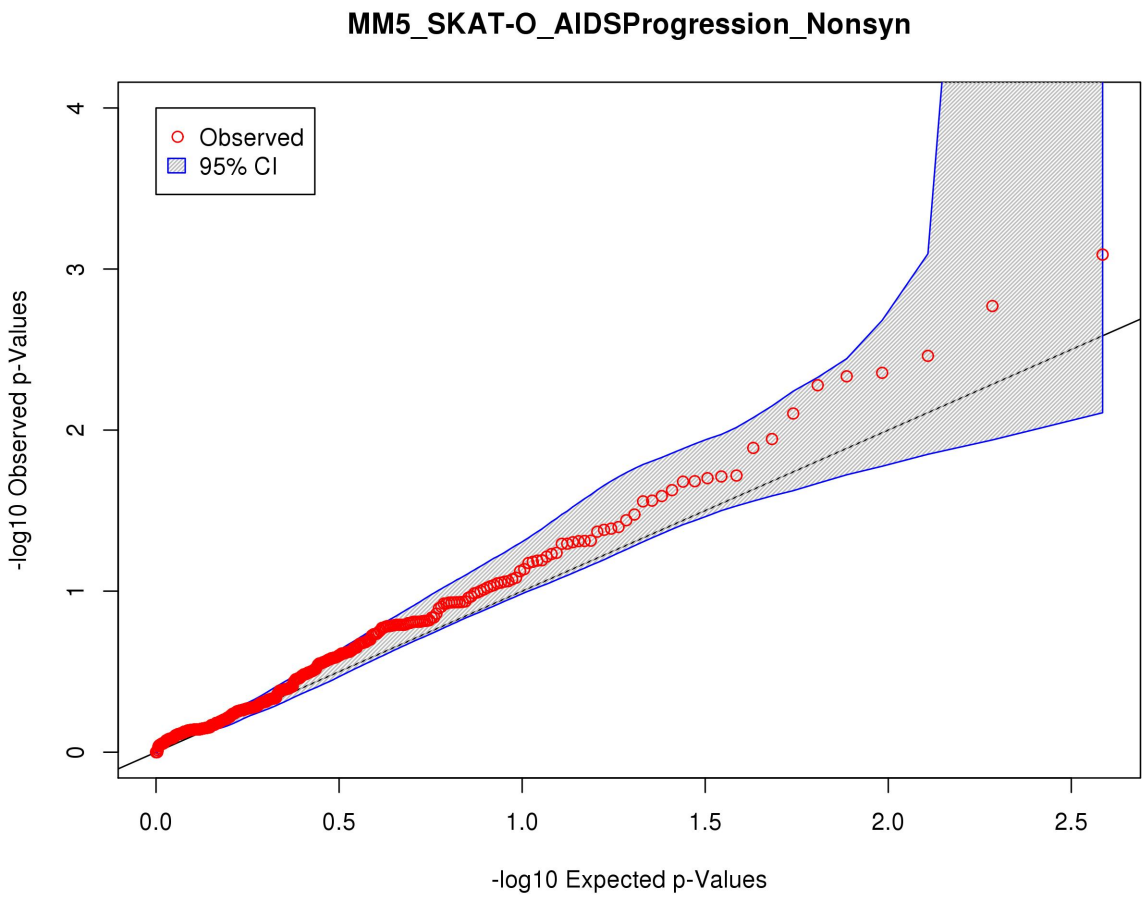

Figure 3.16: **MM5 SKAT-O QQPlot: AIDS-Progression** – Shown is the QQ-Plot for the SKAT-O analyses of the AIDS-Progression phenotype from the MM5 array. AIDS-Progression refers to the analysis of Very Rapid + Rapid AIDS-Progressors vs. Very Slow + Slow AIDS-Progressors. See Methods for descriptions of categories and see Figure 3.3 for remaining details.

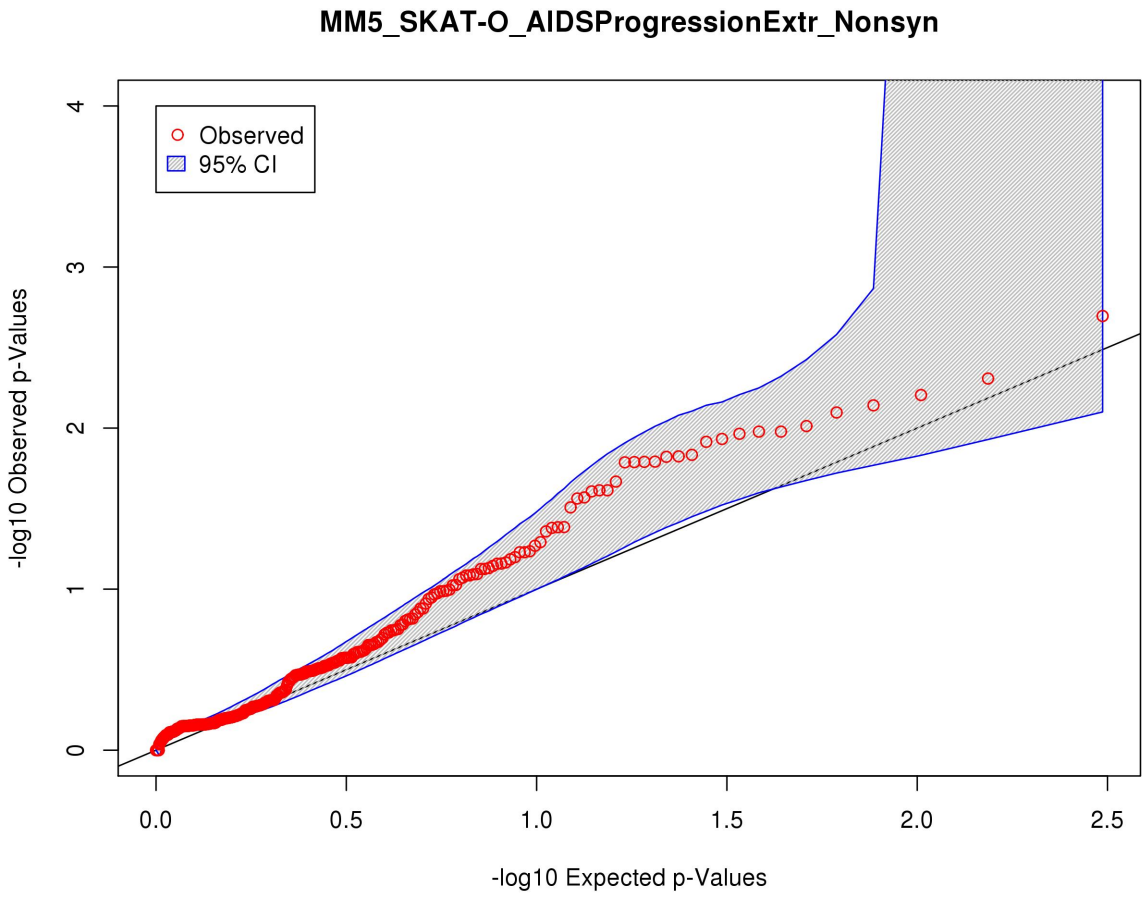

Figure 3.17: **MM5 SKAT-O QQPlot: AIDS-Progression Extr** – Shown is the QQPlot for the SKAT-O analyses of the AIDS-Progression Extr phenotype from the MM5 array. Here extreme ('Extr') refers to only using the Very Rapid AIDS-Progressors and the Very Slow AIDS-Progressors for the SKAT-O analysis. See Methods for descriptions of categories and see Figure 3.3 for remaining details.

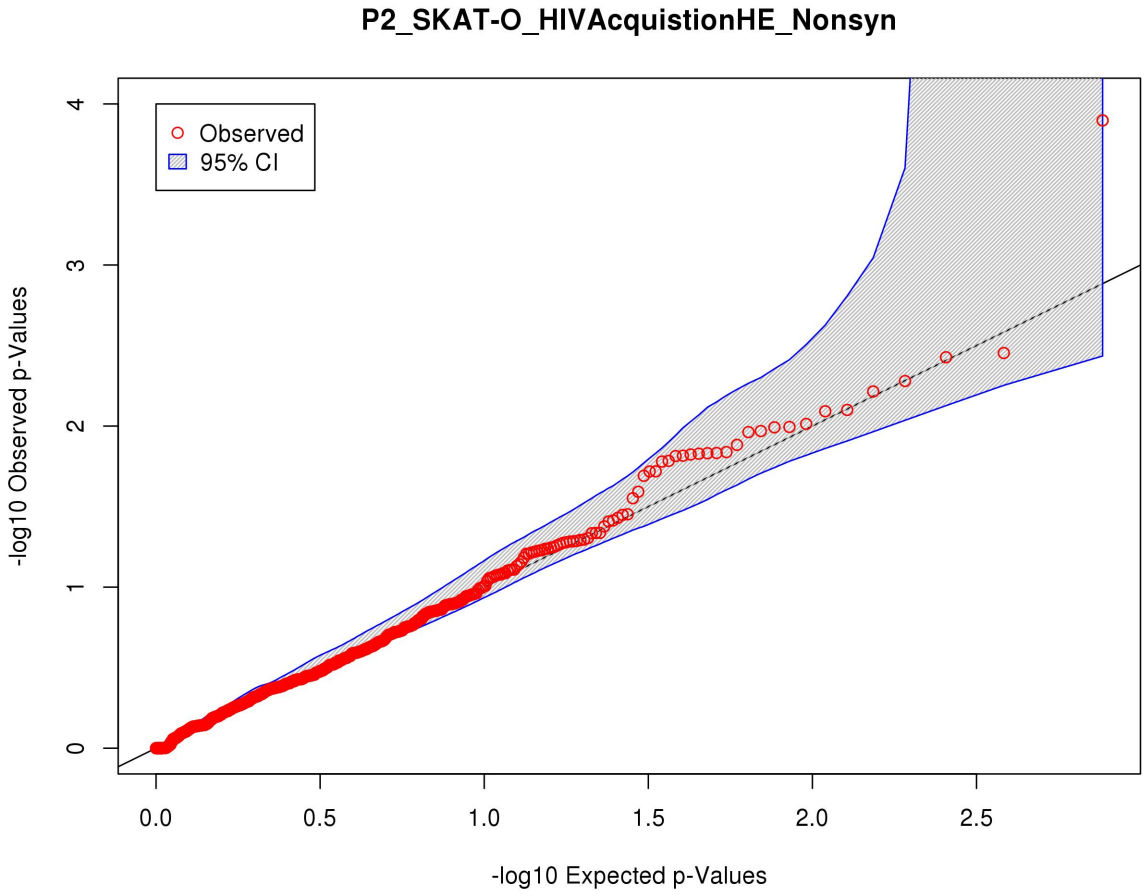

Figure 3.18: **P2 SKAT-O QQPlot: HIV-Acquisition HE** – Shown is the QQPlot for the SKAT-O analyses of the HIV-Acquisition HE phenotype from the P2 array. Here highly-exposed ('HE') seronegative individuals are used as the controls vs. the seropositive cases. See Figure 3.3 for remaining details.

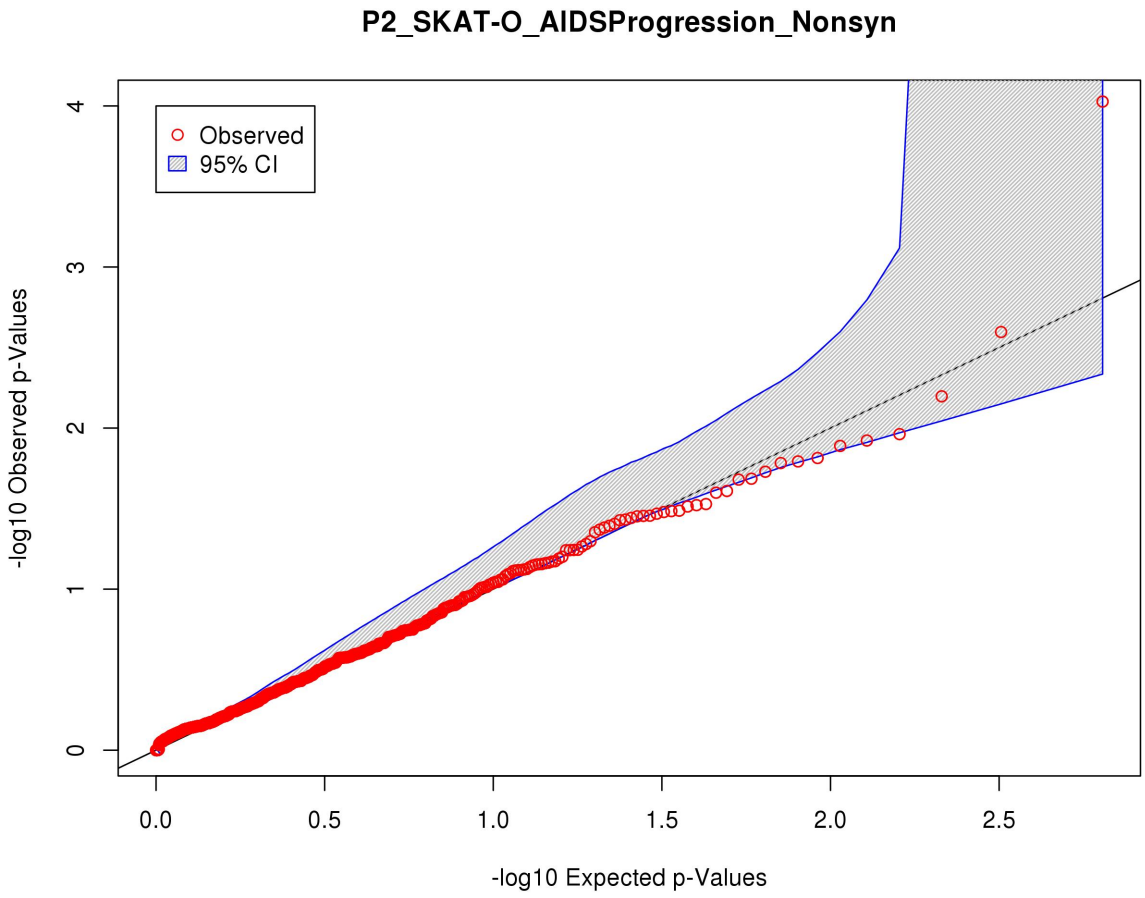

Figure 3.19: **P2 SKAT-O QQPlot: AIDS-Progression** – Shown is the QQPlot for the SKAT-O analyses of the AIDS-Progression phenotype from the P2 array. AIDS-Progression refers to the analysis of Very Rapid + Rapid AIDS-Progressors vs. Very Slow + Slow AIDS-Progressors. See Methods for descriptions of categories and see Figure 3.3 for remaining details.

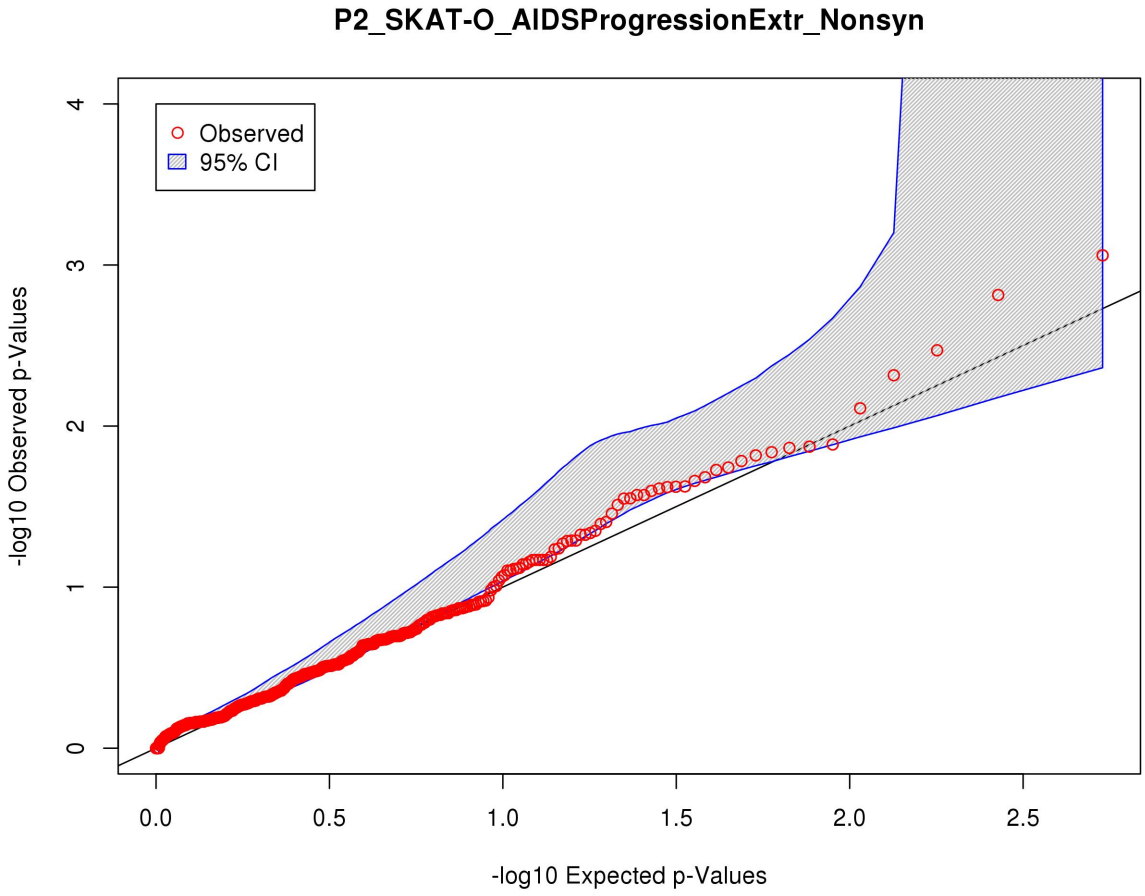

Figure 3.20: **P2 SKAT-O QQPlot: AIDS-Progression Extr** – Shown is the QQPlot for the SKAT-O analyses of the AIDS-Progression Extr phenotype from the P2 array. Here extreme ('Extr') refers to only using the Very Rapid AIDS-Progressors and the Very Slow AIDS-Progressors for the SKAT-O analysis. See Methods for descriptions of categories and see Figure 3.3 for remaining details.

3.11 Supplementary Tables

Ancestry	# Individuals
White non-Hispanic	876
White Hispanic	50
Black non-Hispanic	56
Black Hispanic	1
AmIndian Or AlaskanNative	2
Asian Or PacificIslander	2
Other	1

Table 3.4: **Ancestries of Individuals in MACS Subset** – Table of self-reported ancestries from the individuals included in the MACS subset. Note that only White non-Hispanic individuals were included in the final analysis. Other ancestries are expected to be included in the follow-up genotyping and analysis of the full MACS dataset.

	MM5		P2	
	preQC	**postQC**	**preQC**	**postQC**
SeroNeg	362	351	355	338
HESeroNeg	171	167	164	156
SeroPos	442	424	436	409
RapidVRapid	59	58	57	54
SlowVSlow	92	89	90	81
VRapid	23	23	23	21
VSlow	43	40	42	38

Table 3.5: **White non-Hispanic Pre & Post QC Phenotypes** – Table of phenotype counts for the white non-Hispanic individuals used in the analysis for both the MM5 and P2 arrays. Number of individuals for each phenotype category is shown both pre- and post-QC for both arrays. Post-QC numbers are the final individual counts used for the SKAT-O analyses. SeroNeg refers to seronegative individuals, HESeroNeg refers to highly-exposed seronegative individuals, SeroPos refers to seropositive individuals, RapidVRapid refers to rapid and very rapid AIDS progressor individuals, SlowVSlow refers to slow and very slow AIDS progressor individuals, VRapid refers to very rapid AIDS progressor individuals, and VSlow refers to very slow AIDS progressor individuals.

	MM5			P2		
	All	**SeroNeg**	**SeroPos**	**All**	**SeroNeg**	**SeroPos**
Coverage	23×	22×	23×	15×	15×	15×
Mean GQ	60	59	61	44	44	44
# Exonic	334	334	333	654	654	653
# Non-Synon	123	123	122	257	257	257
# Common	299	299	299	591	591	592
# Rare	35	35	34	62	63	62
# Doubletons	3	3	3	3	3	3
# Singletons	3	2	3	2	2	2

Table 3.6: **MM5 & P2 Variant Summary Metrics** – Shown are various summary metrics from the post-QC set of called variants for both MM5 and P2. Metrics are presented as 'averaged per individual' in the three groupings displayed: all individuals ('All'), seronegative individuals ('SeroNeg'), and seropositive individuals ('SeroPos'). Additionally, metrics shown are produced from focusing on called exonic variants only. The specific summary metrics shown are: mean fold coverage ('Coverage'), mean genotyping quality ('Mean GQ'), mean number of exonic variants ('# Exonic'), mean number of non-synonymous variants ('# Non-Synon'), mean number of common variants (common defined as $\geq$MAF 5%; '# Common'), mean number of rare variants (rare defined as $<$MAF 5%; '# Rare'), mean number of doubletons ('# Doubletons'), and mean number of singletons ('# Singletons').

	MM5			P2		
	q-Val Thresh	**Genes**	**Variants**	**q-Val Thresh**	**Genes**	**Variants**
OrigNonSyn	1	472	3715	1	785	5458
LeftoverNonSyn	1	252	623	1	331	584
AllVarTypes	$\leq$.46	62	6903	$\leq$.52	24	2734
ExACNonSyn	$\leq$.4	40	7267	$\leq$.5	5	756
Blood_eQTLs	NA	213	349	NA	374	636
Total	-	-	18856	-	-	10165

Table 3.7: **Follow-up Genotyping Details: MM5 & P2** – Shown are details of the variants being included for follow-up genotyping of the full MACS cohort. Details are shown for both the MM5 and P2 arrays. For full descriptions of the categories used for inclusion see Methods, but in brief: all the non-synonymous variants used in the original SKAT-O analyses ('OrigNonSyn'), any non-synonymous variants leftover from the original dataset ('LeftoverNonSyn'), variants of any other type other than non-synonymous ('AllVarTypes'), additional non-synonymous SNPs pulled from the ExAC dataset[131] ('ExACNonSyn'), and additional, putative whole blood eQTL SNPs pulled from the GTEX dataset[33]. The columns show for both MM5 and P2 the number of genes and variants that were included for each category. Additionally, where applicable the q-value threshold used to determine which genes would be included for a given category is also shown ('q-Val Thresh'; see Methods). Note that for the GTEX whole blood eQTLs, p-value and basepair distance thresholds were used in lieu of q-values.

CHAPTER 4

BAYESIAN MULTIVARIATE RE-ANALYSIS OF LARGE GENETIC STUDIES IDENTIFIES MANY NOVEL ASSOCIATIONS

Michael C. Turchin[1] and Matthew Stephens[1,2,†]

[1]Department of Human Genetics and [2]Department of Statistics, The University of Chicago

[†]To whom correspondence should be addressed: mstepens@uchicago.edu.

4.1 Abstract

Genome-wide association studies (GWAS) are now a common tool to identify genetic variants that affect traits of interest. To date, the NHGRI GWAS Catalog has over 24,000 SNP-phenotype associations. However, the vast majority of these GWAS are conducted in univariate frameworks, ie when genetic variants are only tested against a single phenotype one at a time. This is in contrast to multivariate frameworks where genetic variants are tested against different combinations of traits simultaneously. Under many biological scenarios, taking into account the context of multiple phenotypes drastically increases power. Additionally, by testing combinations of traits, multivariate frameworks allow researchers to investigate a greater level of biological complexity. Despite these clear advantages, multivariate analyses are seldom implemented. Univariate GWAS already involve a large computational and statistical burden; performing an additional, exponentially greater number of tests is highly deterring. Furthermore, it is often unclear how to properly compare different multivariate models even when they can be efficiently conducted.

Here, we present a framework and R package that alleviates these obstacles – Bayesian multivariate analysis of summary statistics, or bmass. bmass runs solely using univariate GWAS summary statistics. bmass can quickly conduct all possible multivariate analyses for up to 8 phenotypes. And bmass provides Bayes factors for each multivariate analysis, thus allowing models to be directly compared. Running bmass on various publicly available GWAS datasets consistently shows an increase in power up to 40% over univariate approaches while keeping FDRs as low as 15%. bmass

identifies many new significant associations as well as the phenotypic combinations driving these associations, thus providing new levels of biological insight. Overall, bmass is a powerful tool that should further enable researchers to perform multivariate analysis of GWAS.

4.2 Introduction

GWAS have become a common and useful tool for human genetic researchers. Often employed as a first step, GWAS enable researchers to narrow the genome down to regions of potential interest that may be linked to physical traits of interest[174, 227]. In terms of connecting single nucleotide polymorphisms (SNPs) and copy number variations (CNVs) to physical characteristics, over 24,000 such associations have already been identified as 'genome-wide significant' (based on varying thresholds of statistical significance)[143]; additionally, a large, growing number of associations are being discovered between genetic variation and relevant intermediate biological steps such as epigenetic modifications, gene expression, transcript degradation, and other related phenomena[71, 164, 17, 210].

However, these studies often conduct analyses only of the univariate GWAS variety, e.g. they test for association only one trait at a time. This is in contrast to multivariate GWAS approaches, where you would test for association between genetic variation and a combination of traits simultaneously[105, 196, 244]. Despite most data releases of GWAS results being only of the univariate variety, it has been long known and appreciated that multivariate approaches in GWAS have the capacity to increase power to detect genetic association[105, 247, 72]. Additionally, multivariate approaches in GWAS present a natural framework to help interpret genetic associations – by connecting multiple phenotypes to a variant of interest, it both focuses the underlying biological possibilities while also presenting more material to draw conclusions from.

Even with these advantages, multivariate approaches in GWAS are seldom used. There are likely multiple reasons for this, but in this paper we focus on two aspects in particular: difficulties implementing multivariate approaches and difficulties interpreting multivariate results. Often with previous multivariate approaches, users must specify individual models they want to test and then rerun software repeatedly to traverse various possibilities[62, 162]. It can quickly become ungainly and computationally challenging to comprehensively explore many multivariate models, which is a common scenario for most users. Additionally, previous methods often present results in the form of p-values. While this is useful for determining whether there is enough evidence to reject a given null hypothesis, these setups are less conclusive when determining which alternative models are most supported. When exploring various multivariate models, it becomes increasingly important to determine not just which single models are interesting, but which models are more interesting than others.

We address these concerns through the development of bmass – Bayesian multivariate analysis of summary statistics. bmass is an R package that builds on the framework in Stephens 2013[207]. bmass runs on GWAS summary statistics in a standard, simple format, and is able to comprehensively explore all models in our framework quickly given up to 8 different phenotypes. This enhances ease-of-use and is more engaging for users since results for all possible multivariate models are provided. Additionally, presenting results in a Bayesian framework allows users to directly compare models against one another; this in turn eases interpretation of results.

Here we show the results of applying bmass and re-analyzing over 13 different publicly available datasets. For many of these datasets we find additional, novel variants. We also identify multivariate patterns underlying both the previously discovered associations as well as our novel findings. Finally, we highlight specific examples to show how incorporating knowledge from multiple phenotypes can enhance results' interpretation and hypothesis formation.

4.3 Results

4.3.1 Modeling multiple phenotypes and genetic associations in bmass

As mentioned, bmass implements the model from Stephens 2013[207]. For a full description, see both Online Methods and Stephens 2013. In brief, bmass defines phenotypes as belonging to one of three possible categories: **U**nassociated, **D**irectly Associated, or **I**ndirectly Associated. Unassociated traits have no direct connection to, and are independent of, a variant of interest, **g**. Directly associated traits do have a direct connection to **g** and are therefore dependent on **g**. And indirectly associated traits have a connection to **g** but only by mediation through traits in **D**, so therefore are conditionally independent of **g** given **D** (see supplementary figure 4.5 for a graphical structure representing these relationships and Online Methods for more details). bmass then defines a single multivariate model γ as the particular

combination of traits that fall into each of these three categories, eg $\gamma = \{\mathbf{U}, \mathbf{D}, \mathbf{I}\}$. Given a set of $\mathbf{d}$ phenotypes, bmass will define a total of $3^{\mathbf{d}}$ different γ's.

To evaluate these $3^{\mathbf{d}}$ different models, bmass computes a Bayes Factors (BF) for each γ:

$$BF_\gamma := \frac{p_\gamma(\boldsymbol{Y}|\boldsymbol{g})}{p_0(\boldsymbol{Y})} \tag{4.1}$$

where the numerator represents the alternative hypothesis (p_γ) and the denominator represents the null hypothesis (p_0). The denominator is the same for every BF, representing a global null of 'no association' between any of our $\mathbf{d}$ phenotypes and our variant $\mathbf{g}$. And the numerator in our case is each different likelihood corresponding to the γ's, representing different subsets of traits being assigned to $\{\mathbf{U}, \mathbf{D}, \mathbf{I}\}$. bmass calculates these BFs for each variant $\mathbf{g}$ included for analysis, resulting in a $\mathbf{g}$ by $\boldsymbol{\gamma}$ matrix of BFs. To obtain a single summary metric for each variant, bmass employs Bayesian Model Averaging (BMA) and a weighted average:

$$BF_{av} := \Sigma_{\gamma \neq \gamma_0} w_\gamma BF_\gamma \tag{4.2}$$

where the weights w_γ are proportional to the priors on $p(\gamma)$. This final summary metric, one of the main outputs of bmass, can then be viewed as the total, averaged support for association across all possible models versus the global null of no association.

To determine a value of $\mathrm{BF}_{\mathrm{avg}}$ that corresponds with a 'genome-wide significance threshold' per analysis, bmass can take an empirical approach using each study's

previous univariate significant GWAS hits ('PreviousSNPs'). Specifically, bmass first uses these PreviousSNPs, which were determined using dataset-specific GWAS-significant p-value thresholds (see Methods), to train and determine the weights found in BF_{avg}. bmass then uses these weights to calculate BF_{avg} for all variants being analyzed. Lastly, bmass uses the smallest BF_{avg} amongst the PreviousSNPs as its threshold for 'genome-wide significance'; in lieu of a pre-specified value, we use the weakest PreviousSNP as a benchmark. After establishing this value, any SNP that has a BF_{avg} greater than this 'PreviousSNP smallest value' will be considered 'genome-wide significant' in this new multivariate space as well.

In total, given a set of **d** traits measured across **g** variants, bmass can both 1) comprehensively test all multivariate models as defined by our γ's = {**U**, **D**, **I**} and 2) summarize the overall amount of evidence for association vs. no association for each **g**, determining 'genome-wide significance' based on datasets' previous univariate GWAS hits. Additionally, results are presented in the form of BFs. Doing so aids users in interpreting these results in the following ways: 1) for a single model, increasingly larger BFs can be interpreted as increasingly greater support for the alternative hypothesis; this allows results to be viewed not just on a binary scale as 'associated' or 'not associated', but also viewed in terms of 'how strongly associated', and 2) since each BF uses the same null hypothesis, models can be directly compared against one another – eg if model 1 has a larger BF than model 2, we can interpret this as model 1 having greater support than model 2. Lastly, previous comparisons of this framework to other multivariate GWAS methods have shown the framework to either have comparable power or be the most powerful approach in a number of

biologically-relevant scenarios[72, 172].

Some drawbacks of this framework however include the following: 1) We currently assume that all traits have been measured across the same group of individuals. This allows us to estimate the phenotype covariance matrix using only summary statistics, but restricts what phenotypes can be analyzed at one time. 2) The presented strategy requires previous univariate GWAS hits and so cannot be applied to studies that do not have at least a handful of significant hits. 3) The strategy is also computationally intractable for large numbers of phenotypes, due to the exponentially-increasing number of models that are considered (however, bmass can easily deal with the computations for up to about 8 phenotypes).

4.3.2 Many novel loci identified in re-analyzing 13 publicly available GWAS studies

We obtained summary statistics from 13 different publicly available GWAS studies, representing 10 different collections of phenotypes (Table 4.1). Phenotypic collections include blood lipids traits, body morphological traits, red blood cell traits, blood pressure traits, bone density traits, and kidney function-related traits. For three of these phenotypic collections, multiple releases from the source consortiums are included (2 from GlobalLipids[216, 240], 2 from GIANT[128, 203, 87, 241, 138, 198], 2 from HaemgenRBC[225, 11]). We conducted basic QC where possible (see Online Methods).

We ran bmass as described (see Online Methods) using these datasets' univariate summary statistics. Overall, we had a range of success, going from no new loci found in either the MAGIC2010 or SSGAC2016 datasets to over 100 new loci identified in the GIANT2014/5 data. It should be noted that we draw a distinction between 'new hits identified by bmass' and 'truly novel findings' – many studies publish results (eg lists of significant SNPs) based on the meta-analysis of both a discovery dataset and follow-up additional data, such as a replication dataset; however, generally only the discovery set data is publicly released (there are a number of reasons for this often happening, including replication tending to focus on a smaller subset of variants). As a result, we are often only able to run bmass on the discovery dataset; this produces a list of results that are indeed 'new hits' based solely on the discovery data, but in the context of additional follow-up data from the published studies, do not necessarily represent 'novel findings'. In Table 4.2 we delineate how many of our new bmass hits are 'true novel hits' and how many rediscover previous findings that came from the inclusion of additional data.

Unsurprisingly, we see a positive trend between the number of univariate GWAS hits a study produces and the number of new multivariate hits running bmass on that same study produces; while bmass (and multivariate approaches in general) can increase power to detect associations, this effect will always be strongly correlated with the power to detect univariate associations (see Figure 4.1c for how BF_{avg} scales vs. the maximum univariate p-value in GlobalLipids2013). Interestingly, we also see that the variants being identified as 'new multivariate hits' are not necessarily the same variants that would be identified as 'new univariate hits' had we just

relaxed our univariate GWAS p-value threshold (see Figure 4.2 for an example in GlobalLipids2013). In other words, we are not simply increasing power in a generic sense – we are increasing power and detecting variants that are specific to doing a multivariate analysis.

Finally, we have an interesting situation that allows us to partially estimate our true discovery rate. Because we have three consortiums that have multiple releases of their datasets (with the differences generally being larger sample sizes in the later releases), we can ask the question: how many of our bmass hits from the earlier releases overlap with new univariate hits from the later releases? Since both approaches – increasing sample size and using a multivariate approach – are meant to increase power, we might expect there to be a reasonable amount of overlap between these sets of new hits. And indeed using a 50kb window this is what we see, with overlap between results at 50% or better: in GlobalLipids 13 of our 19 bmass hits from the 2010 release overlap with the new set of univariate hits from the 2013 releases, in GIANT 11 of our 15 bmass hits from the 2010 release overlap with the new set of univariate hits from the 2014 and 2015 releases, and in HaemgenRBC 8 of our 15 bmass hits from the 2012 release overlap with the new set of univariate hits from the 2017 release. Additionally, this might be considered a conservative estimate since we may not expect all multivariate hits to quickly reach univariate p-value thresholds given increasing sample sizes. Overall this indicates that bmass is accurate when identifying new significant associations.

4.3.3 Refining association signals within tagged loci via multivariate patterns

On top of indicating new results, looking at multivariate patterns can also help refine association signals in regions with multiple significant SNPs. In our empirical approach, we mask variants within a 1Mb window of a previous univariate GWAS hit; however, even after doing this, some studies produced results that included variants with univariate p-values that technically exceeded the genome-wide significant p-value thresholds used by those studies (Table 4.3). While there are likely multiple reasons why variants that technically surpassed genome-wide thresholds for significance were not ultimately reported by a given study, we hypothesize that a common rationale for doing so is physical proximity to a variant with a stronger signal. For example, more than half of the variants described above are between a 1Mb and 2Mb window of a previous univariate GWAS hit; while many of these variants are not tightly linked to the original univariate hit, it is reasonable to suggest that authors were originally conservative and demurred from reporting these additional, 'nearby' hits. However, by looking at these 'nearby' hits through the prism of multivariate modeling, we often see much clearer pictures of separate patterns of associations between the variants in question.

For example, see rs7515577 and rs12038699, an original GlobalLipids2010 univariate hit and a new GlobalLipids2013 multivariate hit, respectively – rs7515577 reached univariate genome-wide significance in the GlobalLipids2010 dataset and was reported, while rs12038699 reached univariate genome-wide significance in the Glob-

allIpids2013 dataset and was not reported (Table 4.4). rs7515577 and rs12038699 are 549kb apart and have a r^2 of .04. Based on the 2010 results, looking at just the minimum p-value across each phenotype for both SNPs would suggest rs7515577 as the better candidate, and may lead you to mask rs12038699 even in the follow-up 2013 study. However, by looking at the p-values of each phenotype studied simultaneously for both SNPs, you clearly see two different patterns of association (Figure 4.4). We feel that such patterns strongly indicated that rs12038699 should be viewed as a separate genome-wide hit, and in general we often find this is the case for the variants highlighted above.

4.3.4 Identifying pleiotropic patterns of association within a given study

Not only does bmass produce a single summary metric to determine overall association, but bmass also provides BFs for each individual model tested per variant. These can be used to computer posterior probabilities for each model-variant combination by using the priors obtained from the previous univariate GWAS hits and Bayes' rule:

$$p(\gamma|\boldsymbol{Y}, \boldsymbol{g}) \propto p(\gamma)BF_{\gamma} \tag{4.3}$$

We point out two approaches in particular to look at these results. First, we can ask for each new bmass hit identified in a given dataset, 'what model has the highest posterior probability', and then tally the results across each variant. In Table 4.5 we

show an example of such an approach using the GlobalLipids2013 results. Looking at the top results, we find that 38 new variants have the model 'HDL, TG, and TC in **D**, with LDL in **I**' with the greatest posterior probability, 11 variants have the model 'HDL and TC in **D**, with LDL and TG in **I**' with the greatest posterior probability, and 8 variants have the model 'HDL, LDL, and TG in **D**, with TC in **I**' with the greatest posterior probability. Interestingly, we see a number of SNPs contain models where LDL is not assigned to the **D** category; however, it is well appreciated by this point the causal relationship between LDL and other traits such as Cardiovascular Disease[228, 240, 171]. Possibly this could be an artifact of how blood lipids are measured, which may be of importance for downstream analyses.

Second, instead of looking at individual models per variant, we can examine overall support for our three categories **D**, **I**, or **U** per variant. Specifically we can compute the marginal posterior probabilities per variant that each phenotype gets assigned to these categories. As an example, we present results using the GIANT2014/5 data (Figure 4.3). For each of our three categories **D**, **I**, or **U** (Figure 4.3a-c), we show the marginal posterior probability that each SNP belongs to that category for each of the three phenotypes in GIANT2014/5. We also include the ZScores for each variant per phenotype as a comparison (Figure 4.3d). Perhaps unsurprisingly, the vast majority of SNPs have Height assigned to the **D** category. Height is well-appreciated by this point to have a very polygenic architecture[243, 241], suggesting it may play a role even among SNPs where it is not the only association. Additionally, we also see small clusters of SNPs where both Height and BMI are more supported for being in **D**, and SNPs where both Height and WHRadjBMI are more supported for being

in **D**. These clusters possibly represent distinct pathways from SNPs that are solely height-associated.

4.3.5 Examples of novel multivariate discoveries

Amongst our results, we highlight the following novel discoveries as examples of the particular benefits of utilizing a multivariate approach in GWAS.

***rs2862954 & ERLIN1* (GlobalLipids)**

rs2862954 is a synonymous SNP in the ER Lipid Raft Associated 1 (*ERLIN1*) locus, a gene whose protein product is involved in the endoplasmic reticulum-associated degradation of inositol 1,4,5-trisphosphate receptors. It was originally identified as a novel multivariate hit in Stephens 2013[207] based on the analysis of GlobalLipids2010. Interestingly, despite its strong multivariate signal, rs2862954 remained insignificant among the univariate analyses of the updated 2013 GlobalLipids. Examining the p-values across each phenotype between 2010 and 2013 however reveals a trend towards increased significance in HDL and TG (Table 4.6). Additionally, at the time of Stephens2013, *ERLIN1* only had one reference in the literature related to lipid levels, connecting other SNPs in the locus to alanine-aminotransferase (ALT) plasma levels, an important liver enzyme[245]. However, since then, two additional studies have shown *ERLIN1* to be connected to liver function as well. First, a 2013 study of 2,705 European American individuals from the NHLBI Family Heart Study found that *ERLIN1* and rs2862954 specifically were associated both

with ALT and computed tomography-derived measurements of fatty liver (FL)[56]. Interestingly, this study found that separate analyses of ALT and FL failed to reach genome-wide significance, but a combined meta-analysis of both traits did reach significance. Second, a 2015 study of 316 Carribean-Hispanic individuals in New York City looked at 74 candidate nonalcoholic fatty liver disease (NAFLD) SNPs, including rs2862854/*ERLIN1*[49]. Despite the relatively small sample size, the study still found a suggestive protective effect between rs2862954 and NAFLD, as well as other variants in the *ERLIN1-CHUK-CWF19L1* gene cluster.

Thus there is increasing evidence that *ERLIN1* plays an important role in both liver enzyme levels and blood lipid levels. Additionally, this example shows the utility of multivariate approaches when univariate signals do not reach stringent p-value thresholds; in two studies here, a signal of association was only significant after combining effects across phenotypes.

rs11708067 & *ADCY5* (GIANT)

Our second example, rs11708067, is an intronic variant within the adenylate cyclase 5 (*ADCY5*) locus. ADCY5 is an enzyme that helps catalyze the formation of cAMP in response to G protein-coupled receptor signaling, and is a significant hit in our GIANT2014/5 analysis. Its strongest multivariate signal is with Height and BMI in **D**, and WHRadjBMI in **U**, and it has marginal univariate signals of association for both height and BMI (Table 4.6). There exists compelling previous work linking the ADCY5 locus to the phenotypic space inhabited by both height and BMI, which suggests that this variant may represent a causal mutation for this

association signal. Mice homozygous null for ADCY5 show multiple disruptions related to motor function, including impaired coordination, decreased vertical activity, and bradykinesia[102, 18]. Additionally, *de novo* heterozygous mutations in ADCY5 were suggested to be the causative mutations in multiple cases of familial dyskinesia with facial myokymia (FDFM)[31, 30], an autosomal dominant disorder first described in 2001 that presents various motor function dysregulation with early onset[60]. Multiple studies have also shown association signals between rs11708067 and other variants within ADCY5 to Type 2 Diabetes (T2D) and glucose-related measurements[48, 188, 179], as well as fetal growth rate and birth weight[10, 65]. Lastly, rs11708067 has been shown to affect gene expression of ADCY5 in human pancreatic islets[91] (where differential CpG-SNPs and DNA methylation were also found, including at the ADCY5 locus[40]), thus providing a potential mechanistic link between rs11708067 and ADCY5.

Overall, this novel association between rs11708067 and height/BMI furthers the evidence that ADCY5 is a highly pleiotropic locus with impacts on both motor function and metabolic traits. Interestingly, some of this previous work has already suggested ADCY5 as genetic evidence in support of the 'fetal insulin hypothesis' – an alternative explanation for the relationship between low birth-weight and insulin-resistance, stemming from a genetic basis in the offspring in contrast to an impact of the maternal uterine environment[86, 10, 65]. It is possible that the association presented here, between rs11708067 and adult height and BMI, represents perturbations on similar anthropomorphic and metabolic pathways but with less severe outcomes.

4.4 Discussion

Here we present results from reanalyzing publicly available summary GWAS data using a Bayesian multivariate approach, bmass. We show for many studies additional, novel loci are identified by running bmass; we also show that if a study is already well-powered for a univariate analysis, it stands to gain even more from using this multivariate approach. We also show the utility in analyzing multivariate patterns for interpreting results: we show how looking at multiple phenotypes simultaneously can help parse overlapping signals of association, how multivariate models per-SNP allow us to see more broad patterns of biological phenomena (such as genetic architecture in Height), and how multivariate models can aid in developing hypotheses about individual signals of association.

The need for methods and analytical tools that process increasing complexity and dimensionality in human genomic data, such as bmass, is paramount. We are already entering an era with vast sums of publicly available human genetic and phenotypic information, including exciting and promising ventures such as GTEX[33], the UKBioBank[211], and eMERGE[80] However, many of the tools that have been used thus far in human genomics analysis will increasingly be ineffective at leveraging the advantages these multifaceted datasets offer. Much of the promise these types of databases hold is being able to identify broader patterns of association beyond just 'one variant and one phenotype'. While such databases will help evaluate increasingly complex genic models of association, both polygenic[243, 195] and omnigenic[21], they will also help evaluate increasingly complex patterns of pleiotropy among vari-

ants as well[202, 67, 174]. But in order to identify these patterns of association and leverage these datasets, we need to develop and implement methods that explicitly model multivariate connections between variables.

Here we present one possibility to implement multivariate modeling in human GWAS. However, there are multiple ways to begin addressing these kinds of approaches, with other recent work highlighting different, successful strategies. Pickrell et al. 2016[171] take a bivariate approach where they explicitly include direction of effect sizes to address questions of causality; by looking at correlations of effect sizes between traits, and between variants ascertained on a specific trait, they can begin to discern how, and not just whether, two traits are related to one another. Mendelian Randomization (MR) is another approach that aims to discern causality given two specific traits – here a genetic variant is used as an instrument variable to determine whether one trait acts as an intermediate for the other trait[201, 23]. A number of studies have already shown in the context of GWAS how utilizing MR can help interpret both association signals and trait relationships[228, 7, 159]. Explicitly modeling direction of effect and causality, as done in these examples, should be an important priority and goal of ongoing developments in multivariate approaches.

Additionally, analyses shown here with bmass only involve modeling multiple physical traits as outcomes, however publicly available datasets now increasingly include many intermediate phenotypes, including epigenetic data such as DNA methylation, nucleosome positioning, and histone modifications[184], and physical interaction measurements such as results from ChIA-PET and Hi-C studies[69, 45]. In general

the question of how to combine signals from across these distinct, yet presumably linked, measurements, is an ongoing area of work; recent examples show the breadth of datatypes methods are currently attempting to integrate[248, 136, 238]. However, similar to explicitly modeling causality and direction of effects, anticipating outcome variables beyond the historically typical physical traits GWAS are accustomed to is an important priority for ongoing development of multivariate approaches.

We hope researchers find bmass and the underlying framework useful and engaging. While the two main goals of providing such software is to 1) help researchers identify more potential associated loci and 2) allow them to explore the multivariate patterns underlying these signals of association, we also broadly aim to make multivariate approaches more commonplace in human genomics. As outlined here, many of our upcoming challenges in the field will be greatly aided by extending and incorporating multivariate frameworks; human genomics as a field is already rapidly increasing the quantities of data publicly available, and as such we should also work towards scaling the capabilities of our analytical methods and tools as well.

4.5 Online Methods

4.5.1 GWAS Datasets

Below are specific details regarding retrieval and data-processing for each dataset analyzed, including published lists of GWAS hits. Where applicable, sample size

(N), MAF, and p-value thresholds were applied. Additionally, variants which did not contain information for every phenotype, or did not contain the same reference and alternative allele across each phenotype, were dropped. For a handful of studies, external databases were used to retrieve chromosome and basepair information based on rsID# as well.

GlobalLipids2010[216] Original merged, processed, and GWAS-hit annotated summary data from Stephens 2013[207] for HDL, LDL, TG, and TC was downloaded from `https://github.com/stephens999/multivariate` (*dtlesssignif.annot.txt* and *RSS0.txt*).

GlobalLipids2013[240] Summary data for HDL, LDL, TG, and TC was downloaded from `http://csg.sph.umich.edu/abecasis/public/lipids2013/`. A minimum N threshold of 50,000, a MAF threshold of 1%, and a univariate significant GWAS p-value threshold of 5×10^{-8} were used. All variants were oriented to the HDL minor allele. The final merged and QC'd datafile contained 2,004,701 SNPs. rsIDs of published GWAS hits were retrieved for all four phenotypes from `https://www.nature.com/ng/journal/v45/n11/full/ng.2797.html` via Supplementary Tables 2 and 3.

GIANT2010[128, 203, 87] Summary data for Height, BMI, and WHRadjBMI were downloaded from `https://www.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files`. A minimum N threshold of 50,000, a MAF threshold of 1%, and a univariate significant GWAS p-value threshold of 5×10^{-8} were used. Chromosome and basepair position per variant were retrieved from

dbSNP130[194]. All variants were oriented to the Height minor allele. The final merged and QC'ed datafile contained 2,185,686 SNPs. rsIDs of published GWAS hits were retrieved for Height from https://www.nature.com/nature/journal/v467/n7317/full/nature09410.html via Supplementary Table 1, for BMI from https://www.nature.com/ng/journal/v42/n11/full/ng.686.html via Table 1, and for WHRadjBMI from https://www.nature.com/ng/journal/v42/n11/full/ng.685.html via Table 1.

GIANT2014/5[241, 138, 198] Summary data for Height, BMI, and WHRadjBMI were downloaded from https://www.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files. A minimum N threshold of 50,000, a MAF threshold of 1%, and a univariate significant GWAS p-value threshold of 5×10^{-8} were used. Chromosome and basepair position per variant were retrieved from dbSNP130[194]. All variants were oriented to the Height minor allele. The final merged and QC'ed datafile contained 2,340,715 SNPs. rsIDs of published GWAS hits were retrieved for Height from https://www.nature.com/ng/journal/v46/n11/full/ng.3097.html via Supplementary Table 1, for BMI from https://www.nature.com/nature/journal/v518/n7538/full/nature14177.html via Supplementary Tables 1 and 2, and for WHRadjBMI from https://www.nature.com/nature/journal/v518/n7538/full/nature14132.html via Supplementary Table 4.

HaemgenRBC2012[225] Summary data for RBC, PCV, MCV, MCH, MCHC, and Hb were downloaded from the European Genome-Phenome Archive via accession

number EGAS00000000132 https://www.ebi.ac.uk/ega/studies/EGAS00000000132. A minimum N threshold of 10,000, a MAF threshold of 1%, and a univariate significant GWAS p-value threshold of 1×10^{-8} were used. Chromosome and basepair position per variant were retrieved from HapMap release 22[96]. All variants were oriented to the RBC minor allele. The final merged and QC'ed datafile contained 2,327,567 SNPs. rsIDs of published GWAS hits were retrieved for all six phenotypes from https://www.nature.com/nature/journal/v492/n7429/full/nature11677.html via Table 1.

HaemgenRBC2016[11] Summary data for RBC, PCV, MCV, MCH, MCHC, and Hb were shared via personal communication with the authors. A MAF threshold of 1% and a univariate significant GWAS p-value threshold of 8.319×10^{-9} were used. Since sample size was not provided per variant, the following overall study sample sizes were used as proxies per phenotype: 172,952 for RBC, 172,433 for PCV, 173,039 for MCV, 172,332 for MCH, for 172,925 MCHC, and 172,851 for Hb. All variants were oriented to the RBC minor allele. The final merged and QC'ed datafile contained 8,649,095 SNPs. rsIDs of published GWAS hits were determined empirically by using the aforementioned univariate significant GWAS p-value threshold, and greedily pruning SNPs per phenotype file sorted by p-value and using a 1Mb window.

ICBP2011[95, 230] Summary data for SBP, DBP, PP, and MAP were downloaded from dbGaP via accession numberphs000585.v1.p1 (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000585.v1.p1). A minimum

N threshold of 10,000, a MAF threshold of 1%, and a univariate significant GWAS p-value threshold of 5×10^{-8} were used. Chromosome and basepair position per variant were retrieved from HapMap release 21[96]. All variants were oriented to the SBP minor allele. The final merged and QC'ed datafile contained 2,387,851 SNPs. rsIDs of published GWAS hits were retrieved for SBP and DBP from `https://www.nature.com/nature/journal/v478/n7367/full/nature10405.html` via Supplementary Table 5, and for PP and MAP from `https://www.nature.com/ng/journal/v43/n10/full/ng.922.html` via Table 1 and Supplementary Table 2F. Additionally, we gratefully acknowledge the International Consortium for Blood Pressure Genome-Wide Association Studies (Nature. 2011 Sep 11;478(7367):103-9, Nat Genet. 2011 Sep 11;43(10):1005-11) for generating and sharing this data.

MAGIC2010[48] Summary data for FstIns, FstGlu, HOMA_B, and HOMA_IR were downloaded from `https://www.magicinvestigators.org/downloads/`. A MAF threshold of 1% and a univariate significant GWAS p-value threshold of 5×10^{-8} were used. Since sample size was not provided per variant, the overall study sample size of 46,186 was used as a proxy. Chromosome and basepair position per variant were retrieved from HapMap release 22[96]. All variants were oriented to the FstIns minor allele. The final merged and QC'ed datafile contained 2,333,328 SNPs. rsIDs of published GWAS hits were retrieved for all four phenotypes from `https://www.nature.com/ng/journal/v42/n2/full/ng.520.html` via Table 1.

GEFOS2015[246] Summary data for FA, FN, and LS were downloaded from `http://www.gefos.org/?q=content/data-release-2015`. A MAF threshold of .5% and

a univariate significant GWAS p-value threshold of 1.2×10^{-8} were used. Since sample size was not provided per variant, the overall study sample size of 32,965 was used as a proxy. All variants were oriented to the FA minor allele. The final merged and QC'ed datafile contained 8,938,035 SNPs. rsIDs of published GWAS hits were retrieved for all four phenotypes from `https://www.nature.com/nature/journal/v526/n7571/full/nature14878.html` via Supplementary Table 13.

GIS2014[16] Summary data for Iron, Sat, TrnsFrn, and Log10Frtn were shared via personal communication with the authors. A MAF threshold of 1% and a univariate significant GWAS p-value threshold of 5×10^{-8} were used. Since sample size was not provided per variant, the overall study sample size of 48,972 was used as a proxy. All variants were oriented to the Iron minor allele. The final merged and QC'ed datafile contained 1,985,313 SNPs. rsIDs of published GWAS hits were retrieved for all four phenotypes from `https://www.nature.com/articles/ncomms5926/` via Table 1.

SSGAC2016[12] Summary data for NEB_Pooled and AFB_Pooled were downloaded from `https://www.thessgac.org/data`. A MAF threshold of 1% and a univariate significant GWAS p-value threshold of 5×10^{-8} were used. Since sample size was not provided per variant, the following overall study sample sizes were used as proxies per phenotype: 251,151 for NEB_Pooled and 343,072 for AFB_Pooled. All variants were oriented to the NEB_Pooled minor allele. The final merged and QC'ed datafile contained 2,395,561 SNPs. rsIDs of published GWAS hits were retrieved for all four phenotypes from `https://www.nature.com/ng/journal/v48/`

n12/full/ng.3698.html via Table 1.

CKDGen2010/1[119, 19] Summary data for Crea, Cys, CKD, UACR, and MA were downloaded from https://www.nhlbi.nih.gov/research/intramural/researchers/pi/fox-caroline/datasets. A MAF threshold of 1% and a univariate significant GWAS p-value threshold of 5×10^{-8} were used. Since sample size was not provided per variant, the following overall study sample sizes were used as proxies per phenotype: 67,093 for Crea, 20,957 for Cys, 62,237 for CKD, 31,580 for UACR, and 30482 for MA. All variants were oriented to the Crea minor allele. The final merged and QC'ed datafile contained 2,333,498 SNPs. rsIDs of published GWAS hits were retrieved for Crea, Cys, and CKD from https://www.nature.com/ng/journal/v42/n5/full/ng.568.html via Table 2.

EMERGE22015[88] Summary data for ICV, Accumbens, Amygdala, Caudate, Hippocampus, Pallidum, Putamen, and Thalamus were downloaded from http://enigma.ini.usc.edu/research/download-enigma-gwas-results/. A minimum N threshold of 10,000, a MAF threshold of 1% and a univariate significant GWAS p-value threshold of 5×10^{-8} were used. All variants were oriented to the ICV minor allele. The final merged and QC'ed datafile contained 6,271,117 SNPs. rsIDs of published GWAS hits were retrieved for all 8 phenotypes from https://www.nature.com/nature/journal/v520/n7546/full/nature14101.html via Table 1.

4.5.2 Modeling multiple phenotypes

In the multivariate framework bmass extends, the relationship between multiple phenotypes is modeled by use of the three following categories: **U**nassociated, **D**irectly Associated, or **I**ndirectly Associated. Given a variant of interest **g**, we first define these categories as follows: **U** are phenotypes that have no connection with **g**, **D** are phenotypes that have a direct connection with **g**, and **I** are phenotypes that have a connection to **I** only by passing through the phenotypes classified as **D**. To further define these categories, we also claim the following two conditional probability statements: 1) that phenotypes in **U** are independent of **g** and 2) that phenotypes in **I** are conditionally independent of **g** given the phenotypes in **D**. To help visualize this modeling, we show a directed acyclic graph in Supplementary Figure 4.5 which displays the connections between **U**, **D**, and **I**.

Given these three categories and a set of **d** phenotypes, we define a single multivariate model as distinct subsets of **d** falling into any category **U**, **D**, and **I**. We then assign each unique combination of subsets and category-assignments as a $\boldsymbol{\gamma} = \{\mathbf{U}, \mathbf{D}, \text{and } \mathbf{I}\}$. Such that for any set of **d** phenotypes we have $3^{\mathbf{d}}$ possible multivariate models and therefore $3^{\mathbf{d}}$ $\boldsymbol{\gamma}$'s. We also assign a probability distribution $p_{\gamma}(\boldsymbol{Y}|\boldsymbol{g})$ to each $\boldsymbol{\gamma}$, where $\boldsymbol{Y} = \boldsymbol{Y_U}, \boldsymbol{Y_D}, \boldsymbol{Y_I}$, the subsets of phenotypes that are assigned to each category for a given model.

4.5.3 Specifying the form of p_γ and our Bayes Factor

To specify $p_\gamma(\boldsymbol{Y}|\boldsymbol{g})$, we first present its factorized form as the following:

$$p_\gamma(\boldsymbol{Y}|\boldsymbol{g}) = p_\gamma(Y_U)p_\gamma(Y_D|Y_U, \boldsymbol{g})p_\gamma(Y_I|Y_U, Y_D) \tag{4.4}$$

Initially, this may appear to suggest a large number of models need to be defined for every analysis, even for moderate **d**. However, we will next show that it is possible to derive each $\boldsymbol{\gamma}$ from just a subset of models. To set this up, we will present two specific models from which the remaining $\boldsymbol{\gamma}$'s will be constructed from:

1. p_0 – this model specifies all phenotypes in **d** are placed in the **U** category, which can also be thought of as the 'global null'; eg no phenotypes are found as associated with **g**. We assign $p_0(\boldsymbol{Y}|\boldsymbol{g})$ as its corresponding probability distribution.

2. p_1 – this model specifies all phenotypes in **d** are placed in the **D** category, which can also be thought of as the 'full alternative'; eg all phenotypes are found as directly associated with **g**. We assign $p_1(\boldsymbol{Y}|\boldsymbol{g})$ as its corresponding probability distribution.

With these two specific probability distributions defined, we also propose the following two assumptions regarding $p_\gamma(\boldsymbol{Y}|\boldsymbol{g})$:

1. The distributions that do not depend on **g**(the first and the last) are the same as under the global null p_0

2. The distributions that do depend on $\mathbf{g}$(the second) are the same as under the full alternative p_1

Together, these specific models and assumptions allow us to present $p_\gamma(\boldsymbol{Y}|\boldsymbol{g})$ as the following:

$$p_\gamma(\boldsymbol{Y}|\boldsymbol{g}) = p_0(Y_U)p_1(Y_D|Y_U, \boldsymbol{g})p_0(Y_I|Y_U, Y_D) \tag{4.5}$$

with this final form, we are able to fully specify $p_\gamma(\boldsymbol{Y}|\boldsymbol{g})$ in terms of just p_0 and p_1. As a result, we are able to conduct our analysis by specifying a much smaller number of models than previously anticipated, as well as doing so in a manner that is independent of the size of $\mathbf{d}$ as well.

4.5.4 Specifying the form of our Bayes Factor

In this framework we present results in the form of Bayes Factors (BFs). A BF can be thought of as the Bayesian analogue of a likelihood ratio test, where we are comparing the likelihood of an alternative hypothesis to the likelihood of a null hypothesis. Large values for a BF can be viewed as strong support for the alternative hypothesis, whereas small values for a BF can be viewed as strong support for the null hypothesis. In terms of the form for our BF, we can view our starting point as:

$$BF_\gamma := \frac{p_\gamma(\boldsymbol{Y}|\boldsymbol{g})}{p_0(\boldsymbol{Y})} \tag{4.6}$$

where the numerator is our alternative hypothesis, eg the likelihood for a specific γ, $p_\gamma(\boldsymbol{Y}|\boldsymbol{g})$ and the denominator is our global null $p_0(\boldsymbol{Y})$. Expanding the numerator

and denominator to the full form of the $p_\gamma(\boldsymbol{Y}|\boldsymbol{g})$ likelihoods shown in (4.4) gives us:

$$BF_\gamma = \frac{p_\gamma(\boldsymbol{Y}|\boldsymbol{g}) = p_\gamma(Y_U)p_\gamma(Y_D|Y_U,\boldsymbol{g})p_\gamma(Y_I|Y_U,Y_D)}{p_0(\boldsymbol{Y}|\boldsymbol{g}) = p_0(Y_U)p_0(Y_D|Y_U)p_0(Y_I|Y_U,Y_D)} \tag{4.7}$$

and then substituting in our simplified alternative likelihood form (4.5) leads to:

$$BF_\gamma = \frac{p_\gamma(\boldsymbol{Y}|\boldsymbol{g}) = p_0(Y_U)p_1(Y_D|Y_U,\boldsymbol{g})p_0(Y_I|Y_U,Y_D)}{p_0(\boldsymbol{Y}|\boldsymbol{g}) = p_0(Y_U)p_0(Y_D|Y_U)p_0(Y_I|Y_U,Y_D)} \tag{4.8}$$

Finally, with this current form, you will see that the likelihoods for Y_U and Y_I are now identical in the numerator and denominator. This makes intuitive sense since neither Y_U or Y_I depend on $\mathbf{g}$, thus suggesting there should be no difference between them in the alternative and null hypotheses. Canceling these likelihoods out, we are left with:

$$BF_\gamma = \frac{p_1(Y_D|Y_U,\boldsymbol{g})}{p_0(Y_D|Y_U)} \tag{4.9}$$

which becomes the final form of our BF that we will use to test each model γ in our analyses.

4.5.5 Bayesian multivariate regression

To model the relationship between genotype and multiple phenotypic outcomes, we use a Bayesian analogue of multivariate normal regression. To set this up, we first begin with the canonical form of multivariate normal regression:

$$Y = XB + E \tag{4.10}$$

where, in our context, $Y(n \times d)$ is a matrix of d phenotypes measured on each of n individuals; $X(n \times p)$ is a matrix of p covariates measured on the same group of n individuals; $B(p \times d)$ is a matrix of unknown regression coefficients relating the phenotypes to the covariates; and $E(n \times d)$ is a matrix of error terms, whose rows we assume to be independent and identically distributed as $N_d(0, V)$ for some unknown phenotype covariance matrix $V(d \times d)$.

Moving this setup into a Bayesian framework, we now put priors on the unknown components of our model, specifically the matrix of beta coefficients B and the unknown phenotype covariance matrix V. For this step we choose the conjugate priors for (B, V), which are the following:

$$V \sim W^{-1}(\Psi, m) \tag{4.11}$$

$$B|V \sim MN_{p \times d}(0, K^{-1}, V) \tag{4.12}$$

where $W^{-1}(\Psi, m)$ denotes the inverse Wishart distribution with (inverse) scale matrix Ψ and degrees of freedom $m > d - 1$; and $MN_{pxd}(M, V_1, V_2)$ denotes the matrix normal distribution on $p \times d$ matrices, with mean M, and covariance matrices $V_1(p \times p)$ and $V_2(d \times d)$. For further details and considerations regarding this choice in priors, please see Stephens 2013.

Beneficially, use of these priors leads to a marginal likelihood for Y, $p(Y|X, K, \Psi, m)$, that can be computed analytically (see Minka 1999 and Stephens 2013 for greater

detail). The form of this likelihood is:

$$p(Y|X,K,\Psi,m) = \frac{\Gamma_d(n+m)}{\Gamma_d(m)} \frac{|K|^{d/2}}{|X'X+K|^{d/2}} \frac{|\Psi|^{m/2}}{\pi^{nd/2}|RSS(Y|X,K)+\Psi|^{(n+m)/2}} \tag{4.13}$$

where $\Gamma_d(n) = \pi^{d(d-1)/4}\Pi_{i=1}^{d}\Gamma((n+1-i)/2)$ is the multivariate Gamma function, and

$$RSS(Y|X,K) = Y'Y - Y'X(X'X+X)^{-1}X'Y \tag{4.14}$$

is a Bayesian analogue of the residual sums of squares matrix. The distribution of this marginal likelihood is a matrix-t distribution, which we will denote by

$$Y \sim BMVR(X;K,\Psi,m) \tag{4.15}$$

. This emphasizes that it arises from performing a Bayesian MultiVariate Regression of Y on X. With these marginal likelihoods defined, we can now calculate our BFs.

4.5.6 Calculating BFs from univariate GWAS summary information and taking an empirical approach to test the global null

Given a GWAS with **d** phenotypes analyzed independently across p SNPs genome-wide, we determine all relevant BFs for every partition γ for each SNP **g**. We derive these BFs using only the summary statistic information from the GWAS and no raw genotype information. This is possible due to an efficient algorithm that can calculate

the aforementioned marginal likelihoods p_0 and p_1 through the use of three summary covariance matrices, V_{xx}, V_{xy}, and V_{yy}. The definitions of these three statistics, as well as how to calculate them from univariate GWAS summary information, are shown below:

$$V_{xx} := (1/n)g'g \approx 2f(1-f) \qquad \text{(a scalar)} \tag{4.16}$$

$$V_{yx} := (1/n)Y'g \approx \sqrt{2f(1-f)/n}Z \qquad \text{(a } d \text{ vector)} \tag{4.17}$$

$$V_{yy} := (1/n)Y'y \approx (1/p_0)Z_0'Z_0 \qquad \text{(a } d \times d \text{ matrix)} \tag{4.18}$$

where $\boldsymbol{Z}$ are Z-scores derived from univariate GWAS p-values, n is sample size, f is minor allele frequency (MAF), and $\boldsymbol{Z_0}$ are Z-scores from putatively 'null' SNPs, ie variants that have $|Z| < 2$ for all phenotypes analyzed. For derivations and more detail see Stephens 2013. In practice though we are able to calculate the BFs for each γ per SNP through the above quantities and marginal likelihood definitions. To determine genome-wide significance however, we need to formulate a test against the global null across all γ's. We do this by taking a Bayesian Model Averaging (BMA) approach to come up with an overall summary metric per SNP, which we define as BF_{avg} ('avg' for average):

$$BF_{av} := \Sigma_{\gamma \neq \gamma_0} w_\gamma BF_\gamma \tag{4.19}$$

where the weights w_γ are proportional to the prior distribution $p(\gamma)$ and normalized to sum to 1. BF_{avg} represents the cumulative, weighted evidence against the global null across every γ for a given SNP, where each model is weighted on its prior support. To determine these priors, we take an empirical approach – we use the previous univariate GWAS significant SNPs ('PreviousSNPs') from the datasets being analyzed to train our priors. We do this by using an Expectation-Maximization setup; we calculate posterior probabilities per γ per PreviousSNP and then use the average γ posterior across PreviousSNPs as the input prior for the EM's next cycle. Repeating this, we find that the EM consistently converges, giving us the priors needed to weight each γ in our BF_{avg} calculations.

Once we have these BF_{avg}'s per SNP analyzed, we then need to determine a threshold for genome-wide significance. Instead of attempting to find an analogue for a common p-value threshold such as 5×10^{-8}, we once again take an empirical approach using the analyzed datasets' PreviousSNPs. To identify a threshold for genome-wide significance for BF_{avg}, we simply use the minimum BF_{avg} among the PreviousSNPs. The intuition behind this choice is as follows: the PreviousSNPs represent variants that were already found to be statistically genome-wide significant. After transforming these SNPs into this new bmass multivariate space, if other SNPs now have stronger values of BF_{avg} than the weakest PreviousSNP, it stands to reason these other SNPs should be considered genome-wide significant as well. Put another way, we are using the weakest PreviousSNP as a BF_{avg} benchmark to determine genome-wide significance.

In total then this setup allows us to both derive our BFs for every γ per SNP using univariate GWAS summary data, as well as determine genome-wide significance for the overall test of 'association' vs. 'no association' using BF_{avg}. It should be noted that the priors determined in this manner are also used to derive posterior probabilities per γ per SNP as well.

4.6 Acknowledgments

We thank John Novembre, Anna Di Rienzo, and Xin He for continued feedback during the development of this project. We also thank members of the Stephens Lab for helpful input and discussions at various points during the project's implementation. This work was supported by National Institutes of Health Grant R01 HG002585 to MS, and NIH Grants T32 GM007197 and F31 AI118375 to MCT.

4.7 Author Contributions

MS conceived of the original statistical framework and study design. MCT performed the data collection, processing, and analyses. MCT wrote the encapsulating R package bmass. MS supervised the project. MCT and MS wrote the paper.

4.8 Figures

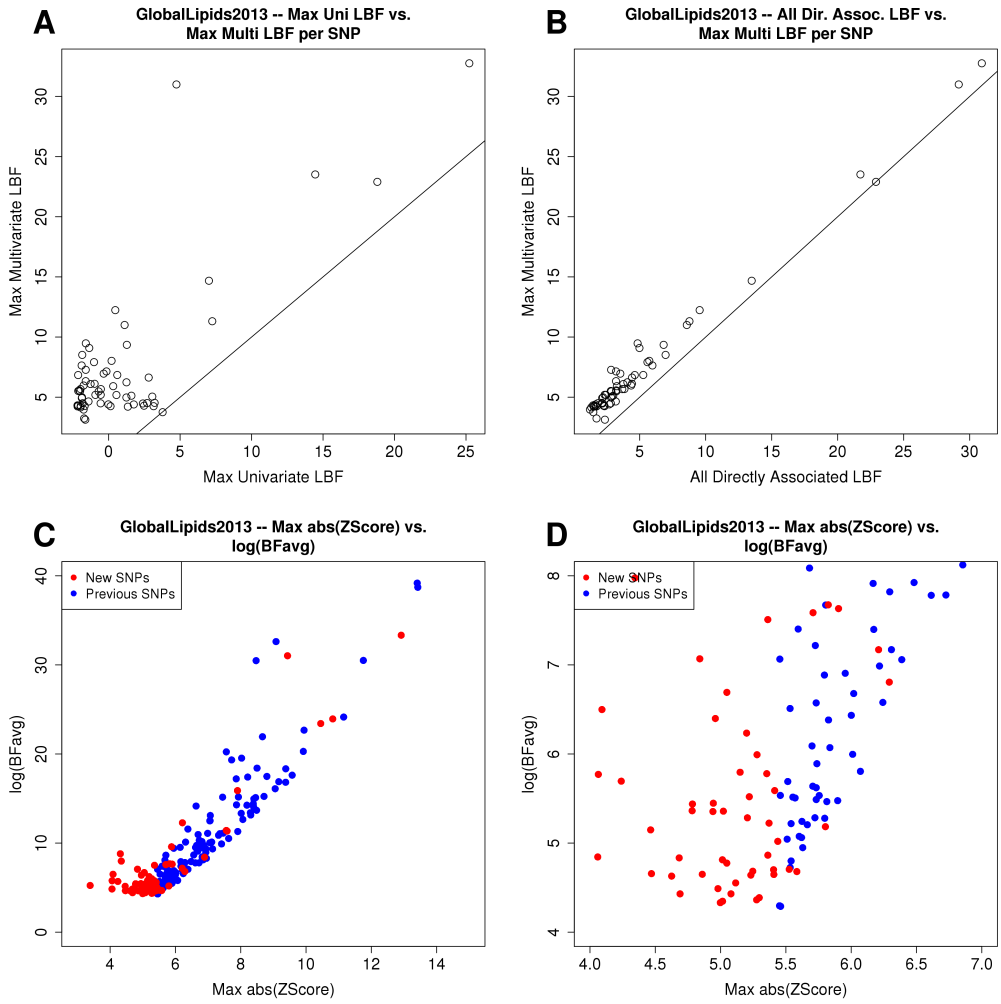

Figure 4.1: **GlobalLipids2013 Model and Metric Comparisons** – Shown are plots comparing some commonly explored models in multivariate settings, as well as a comparison of a common univariate summary statistic versus our method's main multivariate summary statistic. **A)** A plot comparing the max univariate model vs. the max multivariate model in terms of log10 Bayes Factors (LBFs) among the GlobalLipids2010 NewSNPs. **B)** A plot comparing the multivariate model 'all phenotypes in **D**' vs. the max multivariate model in terms of LBFs among the GlobalLipids2010 NewSNPs. **C) & D)** Plots comparing the maximum univariate -log10(pValue) from across all phenotypes analyzed vs. the log10 weighted, average BF (LBFavg) for both the GlobalLipids2010 NewSNPs and PreviousSNPs. **C** shows the plot in full and **D** shows the bottom-left corner zoomed in.

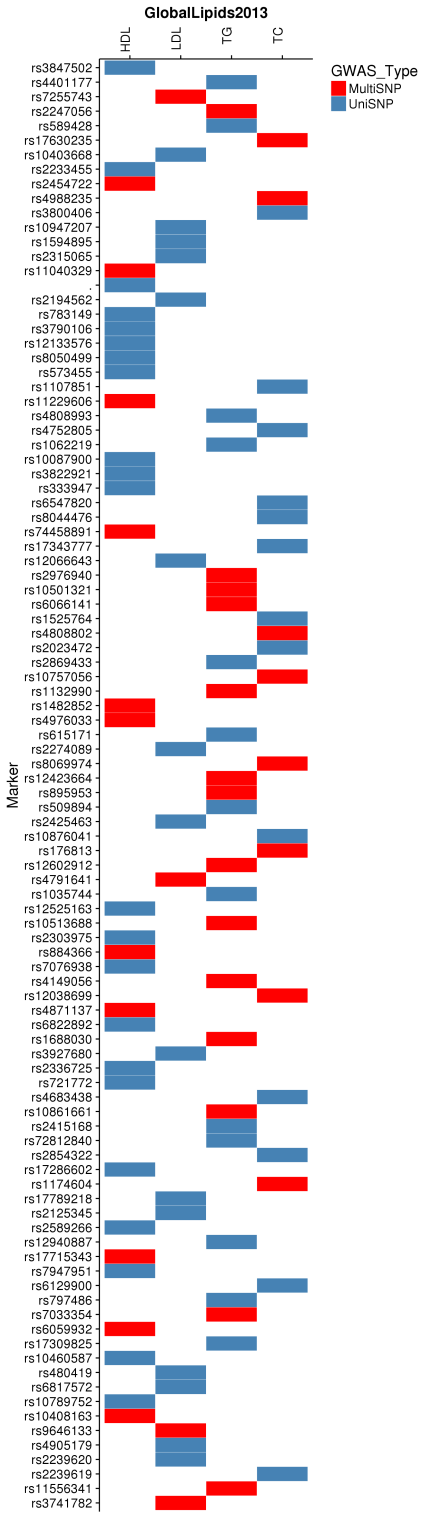

Figure 4.2: **GlobalLipids2013 NewSNPs Significance Ranks**

Figure 4.2 (Cont.): **GlobalLipids2013 NewSNPs Significance Ranks** – Shown here is the distribution of NewSNPs among the ranking of marginally significant, but not GWAS-threshold passing, SNPs. Ranking is based on the minimum p-value from the original univariate analyses of each phenotype and goes in decreasing significance from top down. SNPs are plotted in the column of the phenotype that contained the minimum univariate p-value. Red indicates the SNP was also included as a new hit from the bmass analysis, whereas blue indicates the SNP was only identified by the univariate analysis. Only variants with minimum univariate p-values $<1x10^{-5}$ were included, and variants within 1Mb of a PreviousSNP were removed. Additionally, for display purposes, only the first 150 SNPs are shown.

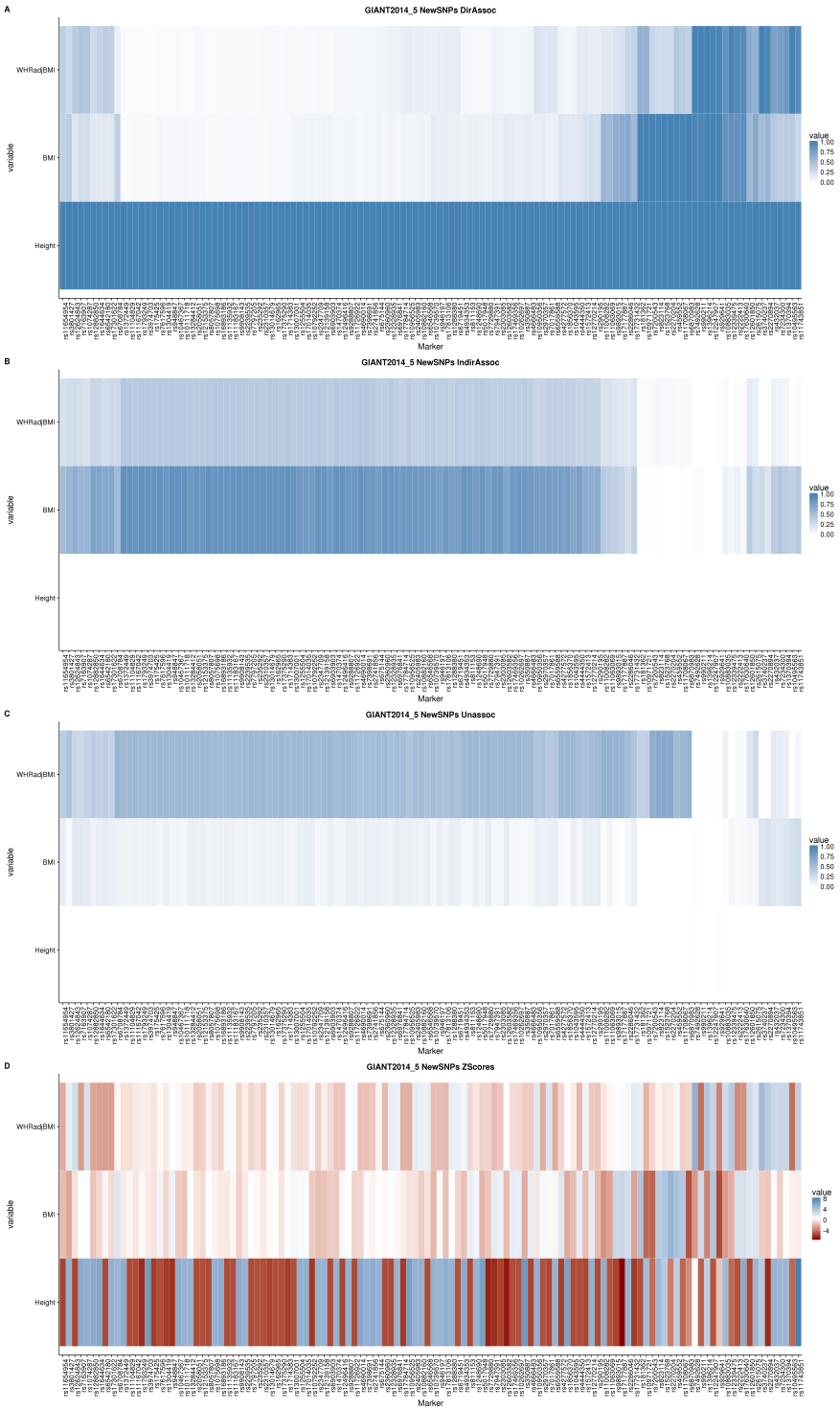

Figure 4.3: **GIANT2014_5 NewSNPs Marginal Posteriors**

Figure 4.3 (Cont.): **GIANT2014_5 NewSNPs Marginal Posteriors** – Shown are heatplots displaying the marginal posterior probabilities of each NewSNP belonging to one of our multivariate categories (**D**, **I**, **U**) for every phenotype analyzed. ZScores for each NewSNP across every phenotype are also displayed. **A)** Marginal Posterior Probabilities of SNP-phenotype combinations being classified as **D**. **B)** Marginal Posterior Probabilities of SNP-phenotype combinations being classified as **I**. **C)** Marginal Posterior Probabilities of SNP-phenotype combinations being classified as **U**. **D)** ZScores for each SNP-phenotype combination. Marginal posterior probabilities for each multivariate category, as well as ZScores, are provided in the supplementary files.

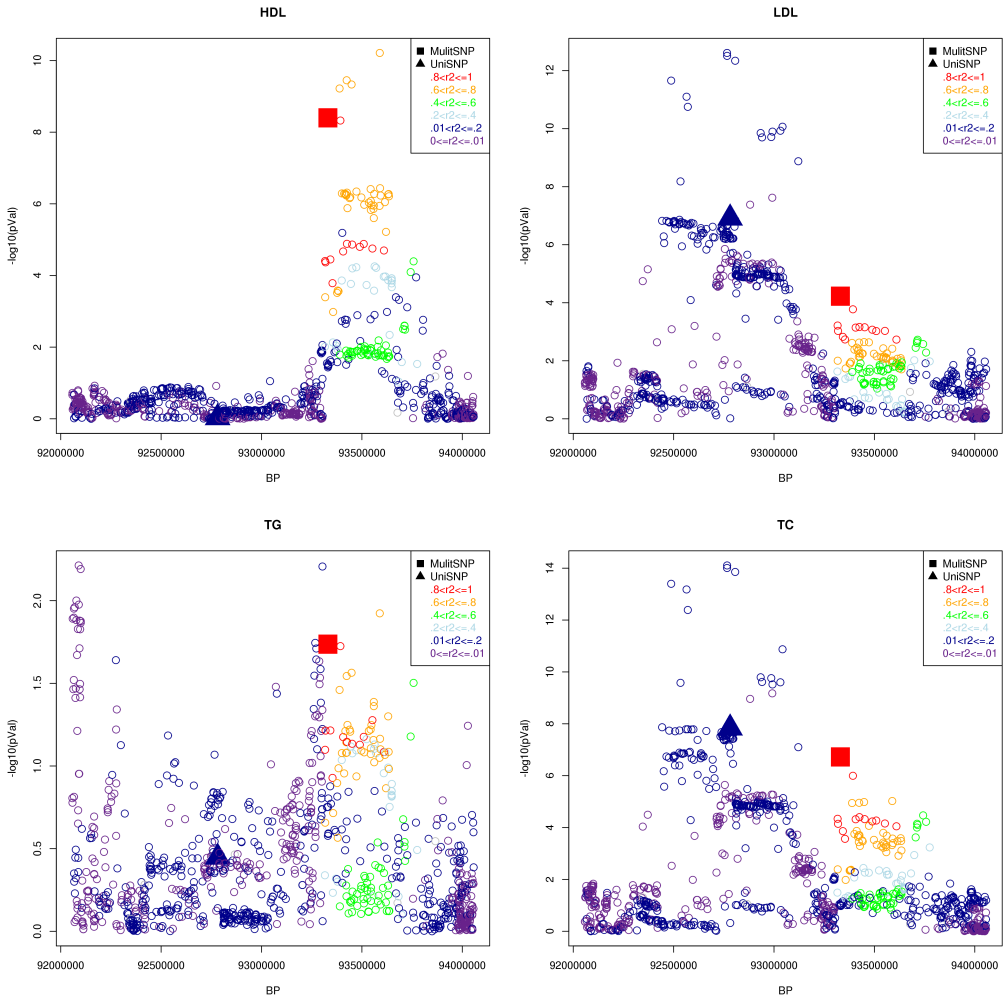

Figure 4.4: **Refining Association Signals – GlobalLipids2013 rs7515577 & rs12038699** – Shown are the -log10 univariate p-values from the GlobalLipids2013 analysis for both the PreviousSNP rs7515577 and the NewSNP rs12038699 across all four phenotypes studied. rs7515577 is represented as a triangle and rs12038699 is represented as a square. Also shown are the -log10 univariate p-values of SNPs in the surrounding 2Mb window of the midpoint between rs7515577 and rs12038699. Color-coding of the SNPs represent the degree of linkage disequilibrium between variants and the NewSNP rs12038699; for color coding details, see legends.

4.9 Tables

Dataset	Release	Max N	Phenotypes
GlobalLipids	2010	95454	LDL, HDL, TC, TG[a]
	2013	188577	LDL, HDL, TC, TG
GIANT	2010	77167	Height, BMI, WHRadjBMI[b]
	2014/5	224459	Height, BMI, WHRadjBMI
HaemgenRBC	2012	135367	RBC, PCV, MCV, MCH, MCHC, Hb[c]
	2016	173480	RBC, PCV, MCV, MCH, MCHC, Hb
ICBP	2011	69395	SBP, DBP, PP, MAP[d]
MAGIC	2010	46186	FstIns, FstGlu, HOMA_B, HOMA_IR[e]
GEFOS	2015	32965	FA, FN, LS[f]
GIS	2014	48972	Iron, Sat, TrnsFrn, Log10Frtn[g]
SSGAC	2016	343072	NEB_Pooled, AFB_Pooled[h]
CKDGen	2010/1	67093	Crea, Cys, CKD, UACR, MA[i]
EMERGE2	2015	30717	ICV, Accumbens, Amygdala, Caudate, Hippocampus, Pallidum, Putamen, Thalamus[j]

Table 4.1: **Dataset Descriptions** – Explanation of phenotypes for each dataset analyzed, as well as the release year and the maximum number of samples (N) per dataset.
a - Low-Density Lipoproteins (LDL), High-Density Lipoproteins (HDL), Total Cholesterol (TC), Total Triglycerides (TG)
b - Waist-Hip Ratio adjusted for BMI (WHRadjBMI)
c - Red Blood Cell Count (RBC), Packed Cell Volume (PCV), Mean Cell Volume (MCV), Mean Cell Haemoglobin (MCH), Mean Cell Haemoglobin Concentration (MCHC), Haemoglobin (Hb)
d - Systolic Blood Pressure (SBP), Diastolic Blood Pressure (DBP), Pulse Pressure (PP), Mean Arterial Pressure (MAP)
e - Fasting Insulin (FstIns), Fasting Glucose (FstGlu), Homeostatic Model Assessment of Beta Cell Function (HOMA_B), HOMA of Insulin Resistance Function (HOMA_IR)
f - Femoral Neck Bone Mineral Density (FN), Forearm BMD (FA), Lumbar Spine BMD (LS)
g - Serum Iron, Serum Transferrin Saturation (Sat), Transferrin (TrnsFrn), Ferritin (Log10Frtn)
h - Number of Children Ever Born (NEB_Pooled), Age at First Birth (AFB_Pooled)
i - Serum Creatine (Crea), Serum Cystatin (Cys), Chronic Kidney Disease (CKD), Urinary Albumin-to-Creatine Ratio (UACR), Microalbuminuria (MA)
j - Intracranial Volume (ICV)

Dataset	Release	Previous Uni Hits	New Multi Hits	BF_{avg} Thresh	Overlap With Next Release
GlobalLipids	2010	100 (+0)	19 (+0)	4.35	13/19
	2013	145 (+2)	63 (+2)	4.23	-
GIANT	2010	128 (+46)	15 (+38)	4.11	11/15
	2014/5	724 (+66)	122 (+40)	4.49	-
HaemgenRBC	2012	63 (+12)	15 (+1)	5.21	8/15
	2016	610 (+0)	60 (+0)	4.27	-
ICBP	2011	22 (+27)	3 (+5)	5.24	-
MAGIC	2010	12 (+3)	0 (+1)	6.90	-
GEFOS	2015	34 (+0)	15 (+0)	5.06	-
GIS	2014	8 (+4)	5 (+0)	7.04	-
SSGAC	2016	9 (+2)	1 (+0)	5.43	-
CKDGen	2010/1	28 (+0)	6 (+0)	4.10	-
EMERGE2	2015	5 (+2)	1 (+2)	7.48	-

Table 4.2: **bmass results** – new multivariate GWAS significant hits based on bmass analysis for all datasets used as well as the number of previous univariate GWAS hits per dataset. The first two columns specify the datasets being analyzed and the years of their release. The third column shows the number of previous univariate GWAS hits based on the discovery dataset. In parentheses is the number of additional univariate GWAS hits identified by that study due to the inclusion of extra data (eg a replication cohort). The fourth column shows the number of novel variants identified by running bmass on the specified dataset. In parentheses are additional variants discovered by running bmass but which are not novel – they are a subset of the variants identified in the original analysis from including extra data. The fifth column shows the minimum PreviousSNP BF_{avg} that was used as the threshold for NewSNP significance. The last column shows for GlobalLipids2010, GIANT2010, and HaemgenRBC2012 the number of novel bmass hits that overlap with the univariate GWAS hits of the dataset's next release; overlap is defined as being within a 50kb window of the univariate GWAS variant.

Dataset	Release	New bmass Hits	Uni pVal	# < Uni pVal	# Nearby Previous Hit[a]
GlobalLipids	2010	19	5×10^{-8}	3	0
	2013	63	5×10^{-8}	18	11
GIANT	2010	15	5×10^{-8}	6	5
	2014/5	122	5×10^{-8}	41	21
HaemgenRBC	2012	15	1×10^{-8}	4	3
	2016	60	8.31×10^{-9}	0	0
ICBP	2011	3	5×10^{-8}	2	1
GEFOS	2015	15	1.2×10^{-8}	4	1
GIS	2014	5	5×10^{-8}	1	1
CKDGen	2010/1	6	5×10^{-8}	0	0
EMERGE2	2015	1	7.1×10^{-9}	0	0

Table 4.3: **New bmass hits and univariate GWAS p-value thresholds** – List of the new bmass hits per dataset that have a univariate p-value below the original GWAS threshold (eg 5×10^{-8}) specified by that dataset. Shown under # Nearby Previous Hit are which among these variants are also nearby a previous univariate GWAS hit, where nearby is defined as between a 1 and 2Mb window.

SNP	Pheno.	Direct.	2010	2013
rs7515577				
(Previous)	HDL	+	9.81E-01	9.29E-01
	LDL	-	1.51E-07	1.21E-07
	TG	-	1.80E-01	3.57E-01
	TC	-	2.78E-08	1.47E-08
rs12038699				
(New)	HDL	+	4.22E-05	3.98E-09
	LDL	+	1.06E-03	5.95E-05
	TG	+	8.51E-02	1.83E-02
	TC	-	7.12E-05	1.90E-07

Table 4.4: **rs7515577 & rs12038699 p-values** – p-values from both the 2010 and 2013 GlobalLipids datasets for SNPs rs7515577 and rs12038699. It can be seen under the 2010 release that rs7515577 has a univariate p-value that crosses the 5×10^{-8} threshold (TC), whereas rs12038699 does not. Since rs12038699 is nearby rs7515577, it is likely to effectively get masked for future analyses; however looking at the 2013 data it is clearly seen that rs12038699 not only has a lower minimum univariate p-value, but also has a distinct multivariate p-value pattern as compared to rs7515577. Both of these signals support the notion that rs7515577 should be viewed as a separate GWAS hit for GlobalLipids2013.

	HDL_LDL_TG_TC	n	MeanPosterior	OriginalPrior
NewSNPs				
	1_2_1_1	38	0.622	0.327
	1_2_2_1	11	0.428	0.138
	1_1_1_2	8	0.577	0.117
	1_1_2_0	3	0.875	0.011
	2_1_2_2	1	0.26	0.057
PreviousSNPs				
	1_2_1_1	59	0.597	0.327
	1_2_2_1	24	0.469	0.138
	2_1_2_2	14	0.498	0.057
	1_1_1_1	10	0.766	0.078
	1_1_1_2	8	0.613	0.117

Table 4.5: **GlobalLipids2013 Top Multivariate Models** – List of top multivariate models that most frequently have the highest posterior probability per SNP. Top 5 models (at most) are shown for NewSNPs and PreviousSNPs analyzed. Phenotype ordering is shown in the header, where 0, 1, and 2 refer to the multivariate categories of **U**nassociated, **D**irectly Associated, and **I**ndirectly Associated. n is the number of SNPs that show the specified model as having the largest posterior probability, with MeanPosterior displaying the average posterior probability of the given model across the n SNPs, and OriginalPrior showing the prior established for the given model from training on the PreviousSNPs.

SNP	Pheno.	Direct.	2010	2014/5
rs11708067				
	Height	+	3.85E-03	1.00E-05
	BMI	-	4.66E-02	6.23E-05
	WHRadjBMI	-	1.70E-01	3.50E-01

Table 4.6: **rs11708067 p-values** – p-values from both the 2010 and 2014_5 GIANT datasets for SNP rs11708067. While in neither the 2010 or 2013 releases rs11708067 reaches univariate GWAS significance (5×10^{-8} for both GIANT releases), a trend towards lower p-values can be seen in both Height and BMI from 2010 to 2014_5. Additionally, both Height and BMI in 2014_5 can be considered 'Marginally Significant', increasing our prior expectations that this SNP may eventually be shown as a true hit.

4.10 Supplementary Figures

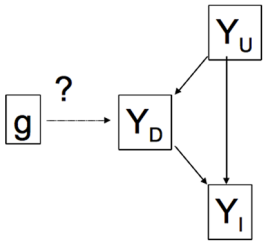

Figure 4.5: **Graphical Model of Multivariate Categories** – Shown here is a Directed Acyclic Graphical (DAG) model of our multivariate categories in the context of our vector of phenotypes $\mathbf{Y}$ (eg $\mathbf{Y} = \{\mathbf{Y_U}, \mathbf{Y_D}, \mathbf{Y_I}\}$) and their connections with the variant of interest $\mathbf{g}$. The relationships described in-text can be seen here. $\mathbf{Y_U}$, our unassociated phenotypes, have no connection with $\mathbf{g}$. $\mathbf{Y_D}$, our directly associated phenotypes, have a direct connection with $\mathbf{g}$. And $\mathbf{Y_I}$, our indirectly associated phenotypes, have a connection with $\mathbf{g}$ only by going through $\mathbf{Y_D}$ first. Note that if $\mathbf{Y_D}$ were not observed, $\mathbf{Y_I}$ would appear as a direct connection.

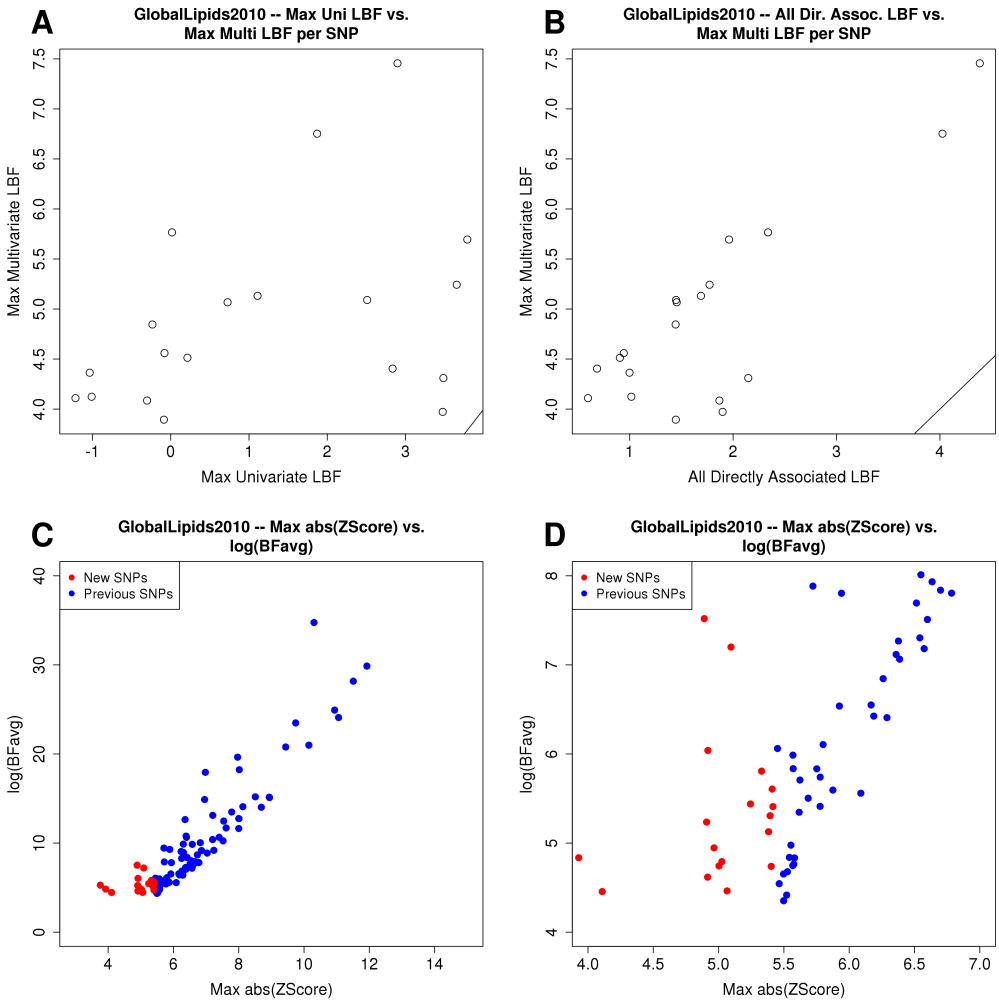

Figure 4.6: **GlobalLipids2010 Model and Metric Comparisons** – Comparison of univariate and multivariate models and summary metrics for GlobalLipids2010 NewSNPs and PreviousSNPs. See GlobalLipids20134.1 for full description.

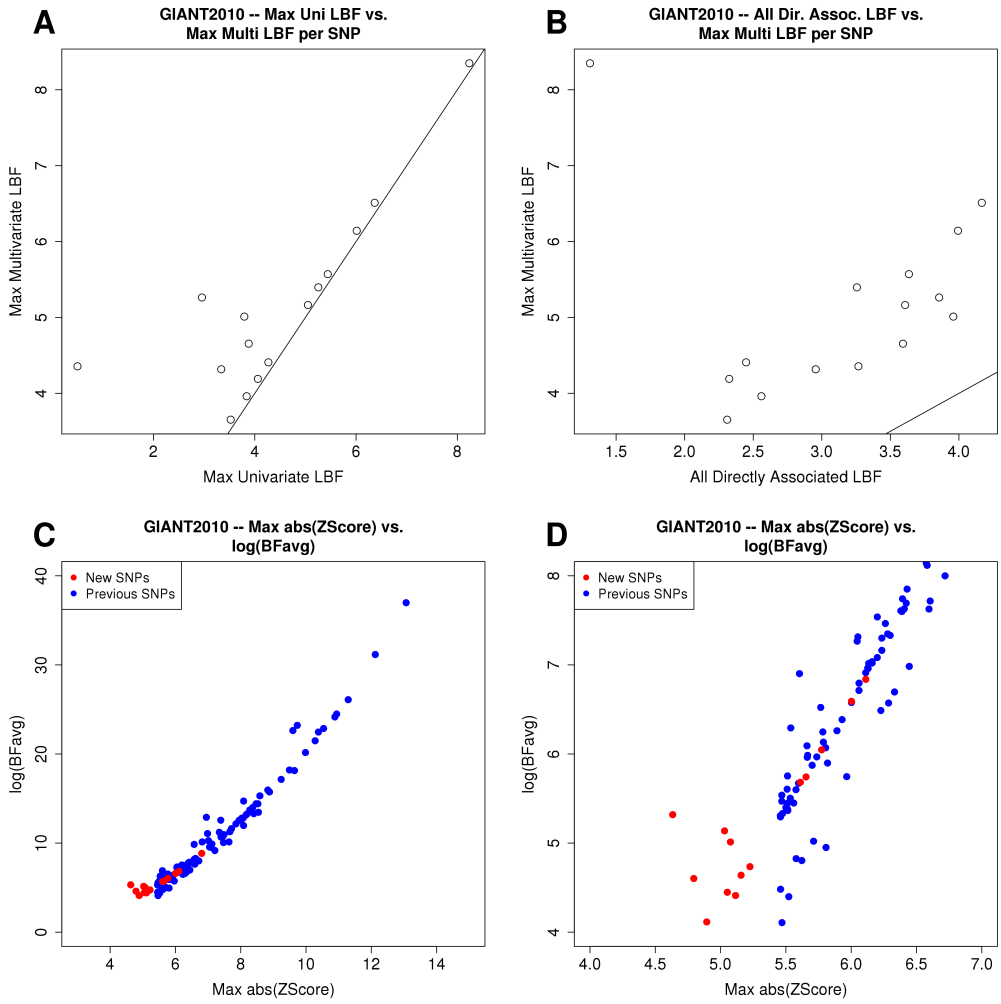

Figure 4.7: **GIANT2010 Model and Metric Comparisons** – Comparison of univariate and multivariate models and summary metrics for GIANT2010 NewSNPs and PreviousSNPs. See GlobalLipids2013 (4.1) for full description.

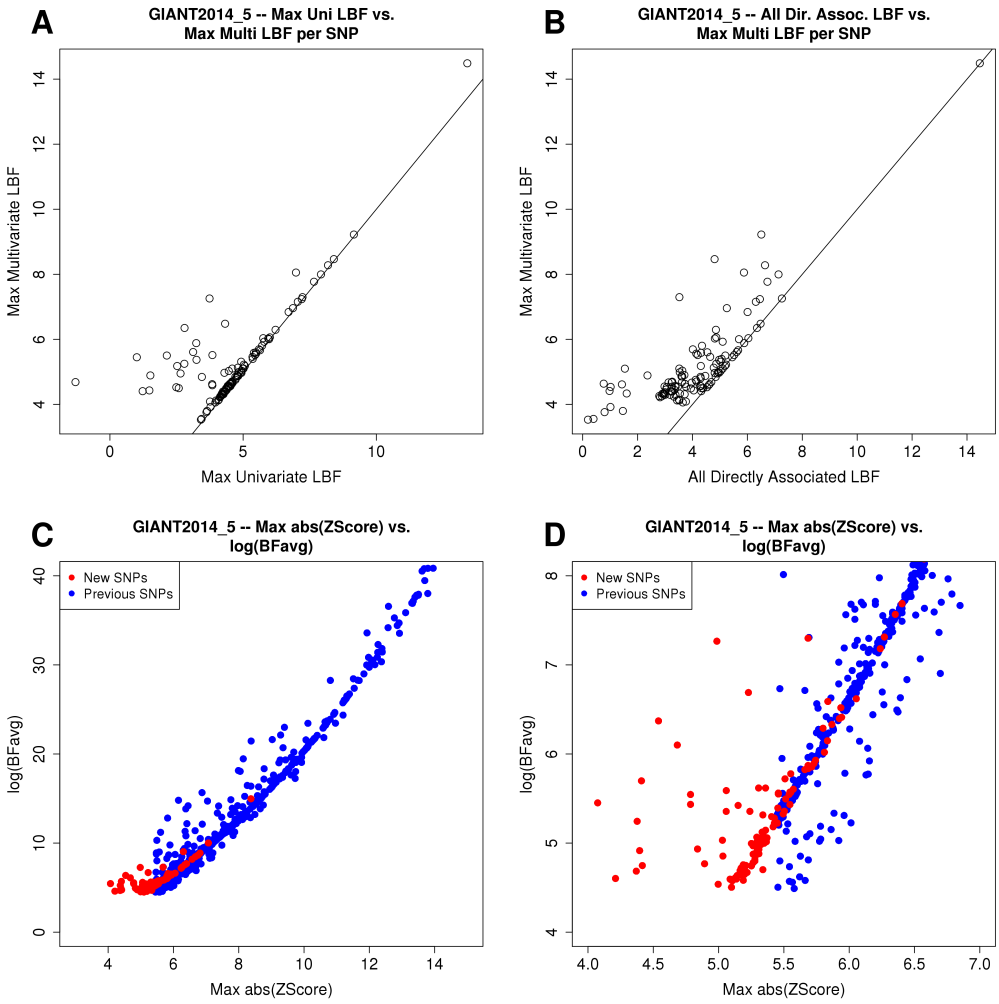

Figure 4.8: **GIANT2014_5 Model and Metric Comparisons** – Comparison of univariate and multivariate models and summary metrics for GIANT2014_5 NewSNPs and PreviousSNPs. See GlobalLipids2013 (4.1) for full description.

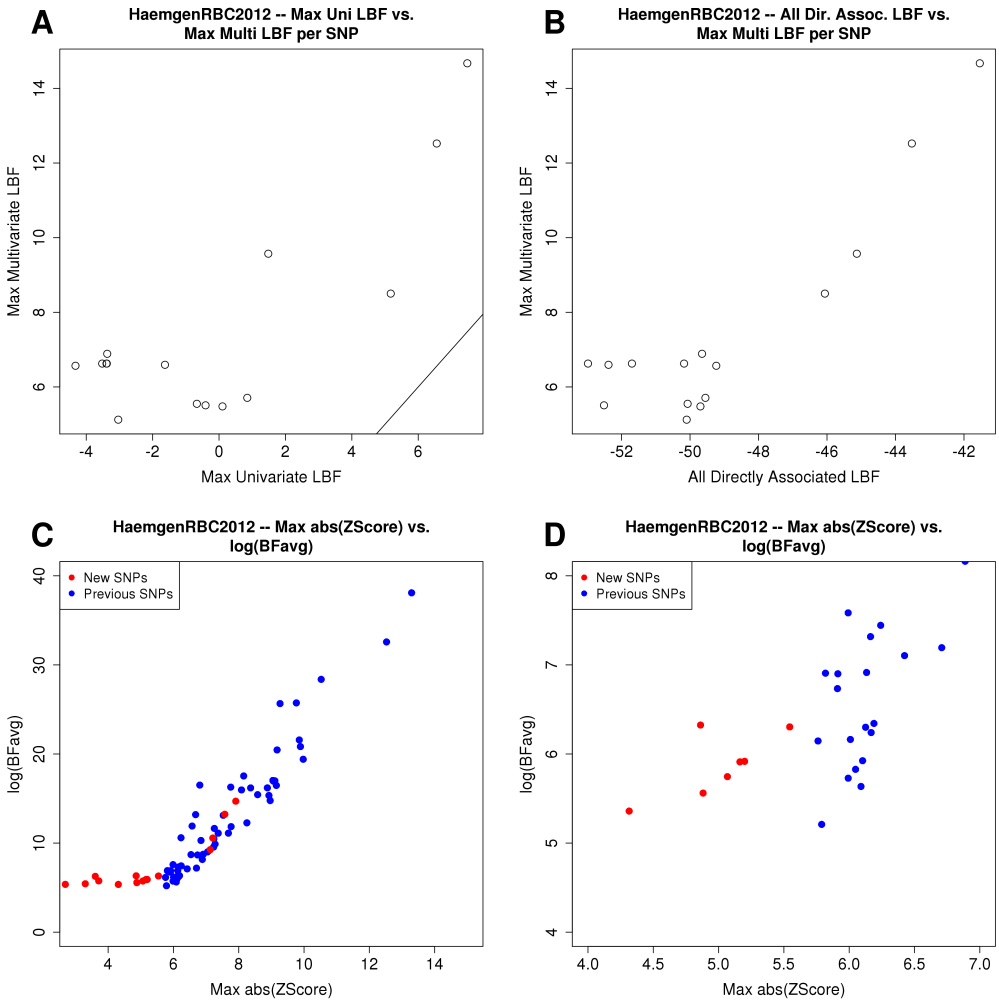

Figure 4.9: **HaemgenRBC2012 Model and Metric Comparisons** – Comparison of univariate and multivariate models and summary metrics for HaemgenRBC2012 NewSNPs and PreviousSNPs. See GlobalLipids2013 (4.1) for full description.

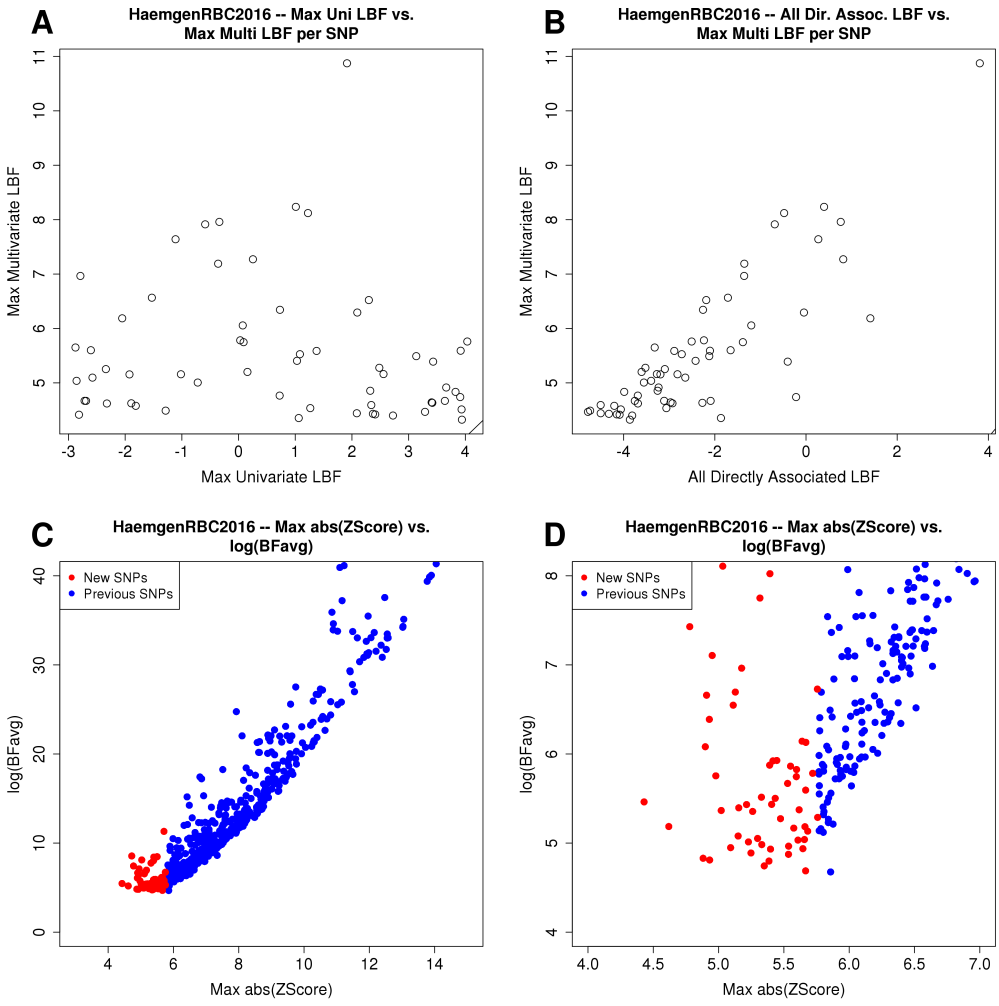

Figure 4.10: **HaemgenRBC2016 Model and Metric Comparisons** – Comparison of univariate and multivariate models and summary metrics for HaemgenRBC2016 NewSNPs and PreviousSNPs. See GlobalLipids2013 (4.1) for full description.

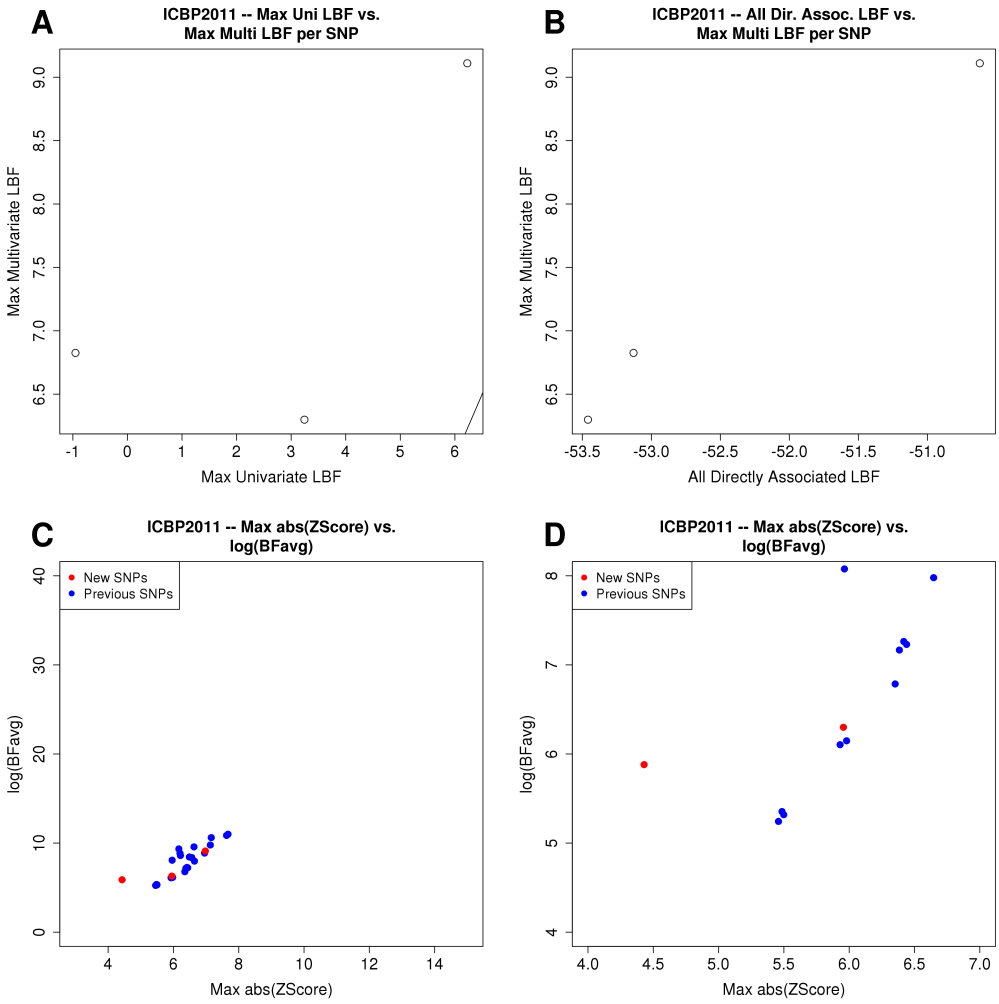

Figure 4.11: **ICBP2011 Model and Metric Comparisons** – Comparison of univariate and multivariate models and summary metrics for ICBP2011 NewSNPs and PreviousSNPs. See GlobalLipids2013 (4.1) for full description.

MAGIC2010 Model and Metric Comparisons – No figure is presented for MAGIC2010 as there are no NewSNPs.

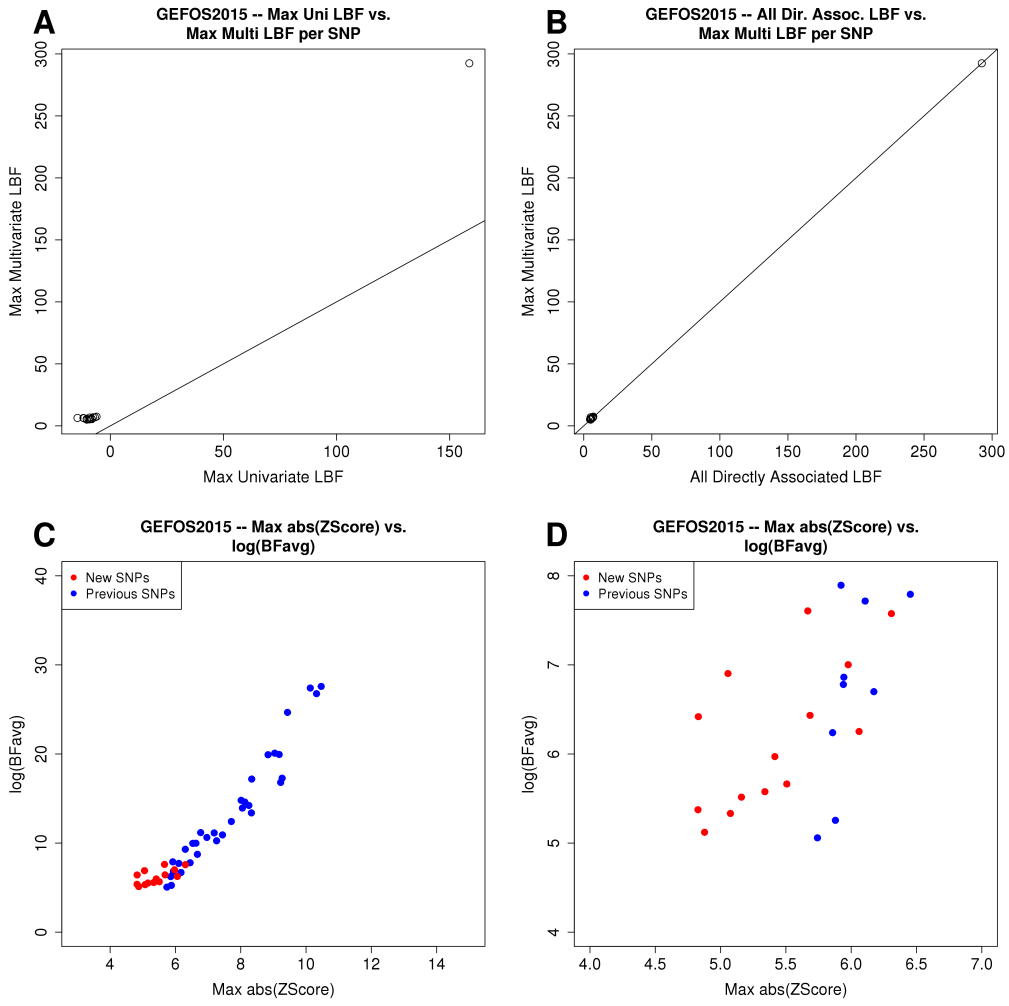

Figure 4.12: **GEFOS2015 Model and Metric Comparisons** – Comparison of univariate and multivariate models and summary metrics for GEFOS2015 NewSNPs and PreviousSNPs. See GlobalLipids2013 (4.1) for full description.

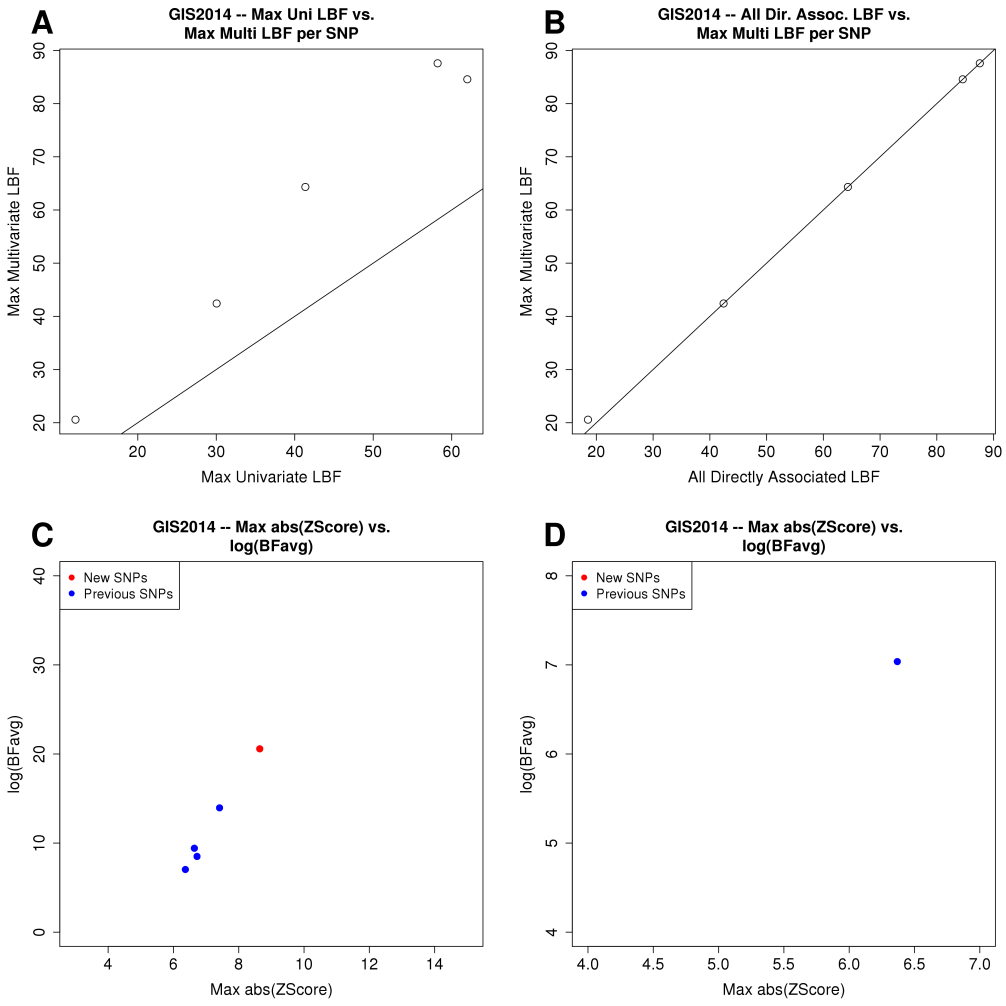

Figure 4.13: **GIS2014 Model and Metric Comparisons** – Comparison of univariate and multivariate models and summary metrics for GIS2014 NewSNPs and PreviousSNPs. See GlobalLipids2013 (4.1) for full description.

SSGAC2016 Model and Metric Comparisons – No figure is presented for SSGAC2016 as there is only a single NewSNP.

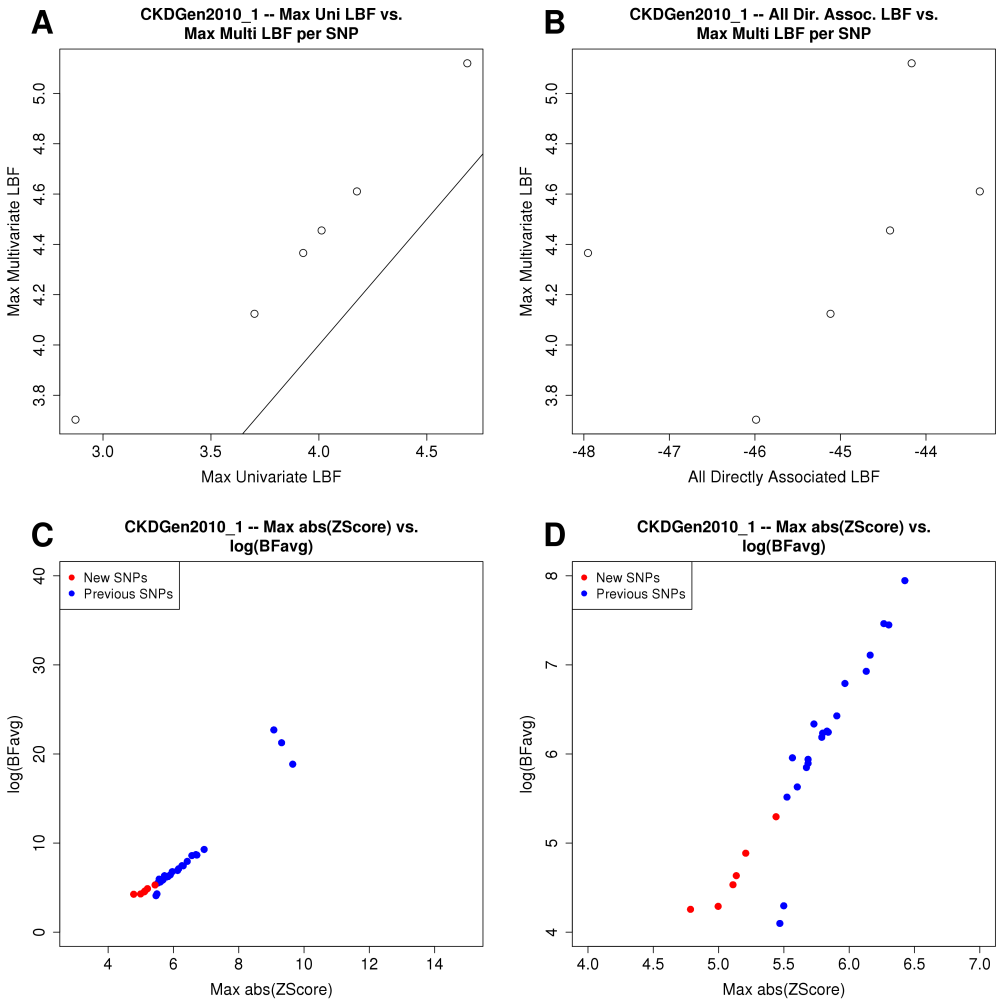

Figure 4.14: **CKDGen2010_1 Model and Metric Comparisons** – Comparison of univariate and multivariate models and summary metrics for CKDGen2010_1 NewSNPs and PreviousSNPs. See GlobalLipids2013 (4.1) for full description.

EMERGE22015 Model and Metric Comparisons – No figure is presented for EMERGE22015 as there is only a single NewSNP.

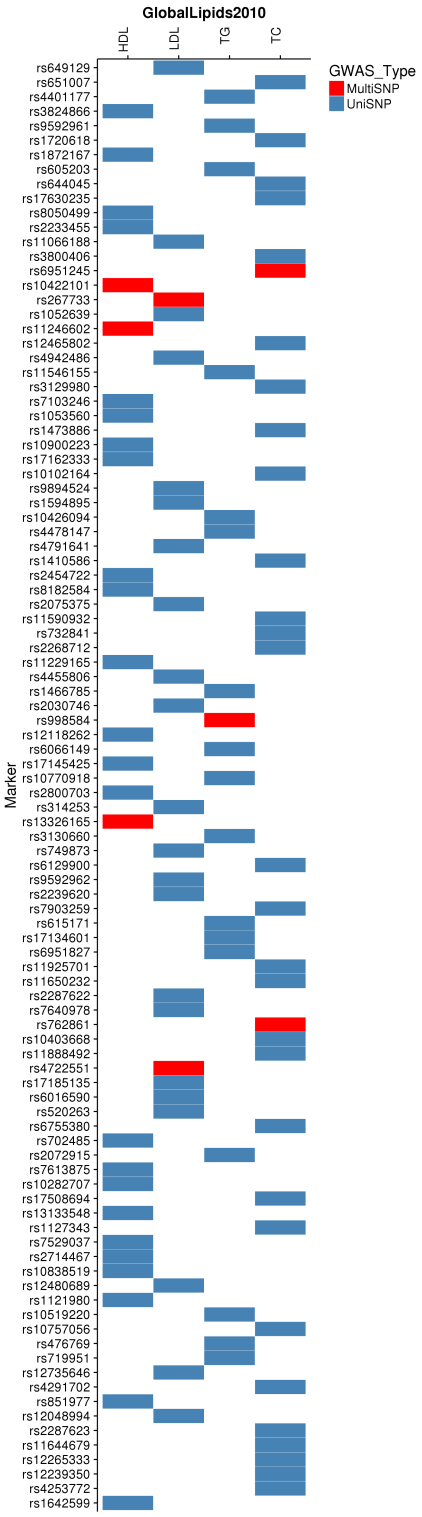

Figure 4.15: **NewSNPs Significance Ranks**

Figure 4.15 (Cont.): **NewSNPs Significance Ranks** – Distribution of NewSNPs among the ranking of marginally significant SNPs for GlobalLipids2010. See GlobalLipids20134.2 for full description.

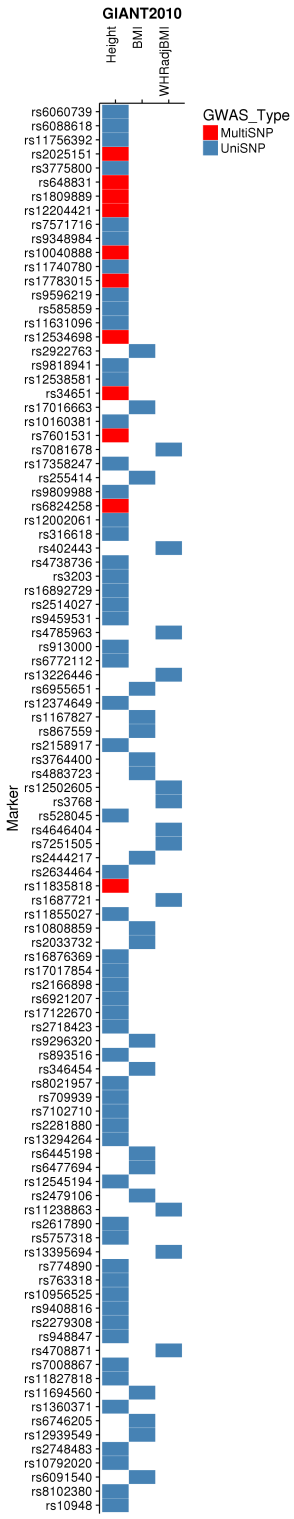

Figure 4.16: **GIANT2010 NewSNPs Significance Ranks**

Figure 4.16 (Cont.): **GIANT2010 NewSNPs Significance Ranks** – Distribution of NewSNPs among the ranking of marginally significant SNPs for GIANT2010. See GlobalLipids2013 (4.2) for full description.

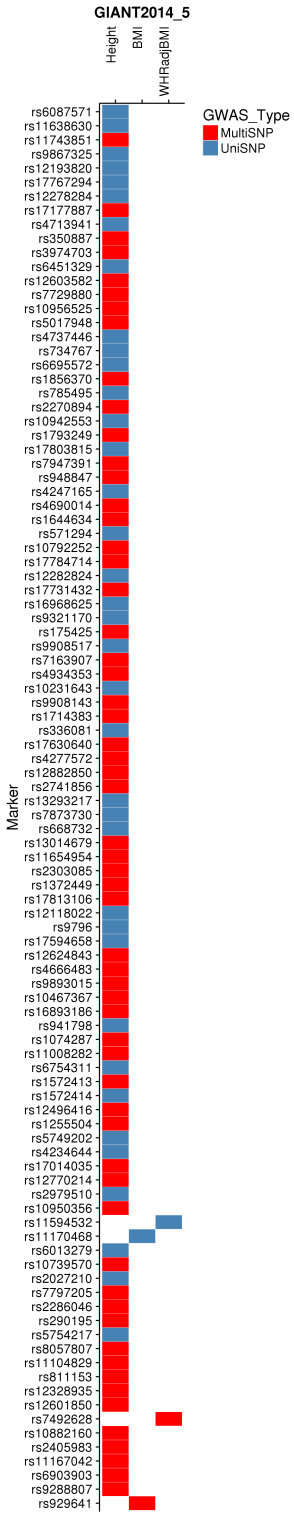

Figure 4.17: **GIANT2014_5 NewSNPs Significance Ranks**

Figure 4.17 (Cont.): **GIANT2014_5 NewSNPs Significance Ranks** – Distribution of NewSNPs among the ranking of marginally significant SNPs for GIANT2014_5. See GlobalLipids2013 (4.2) for full description.

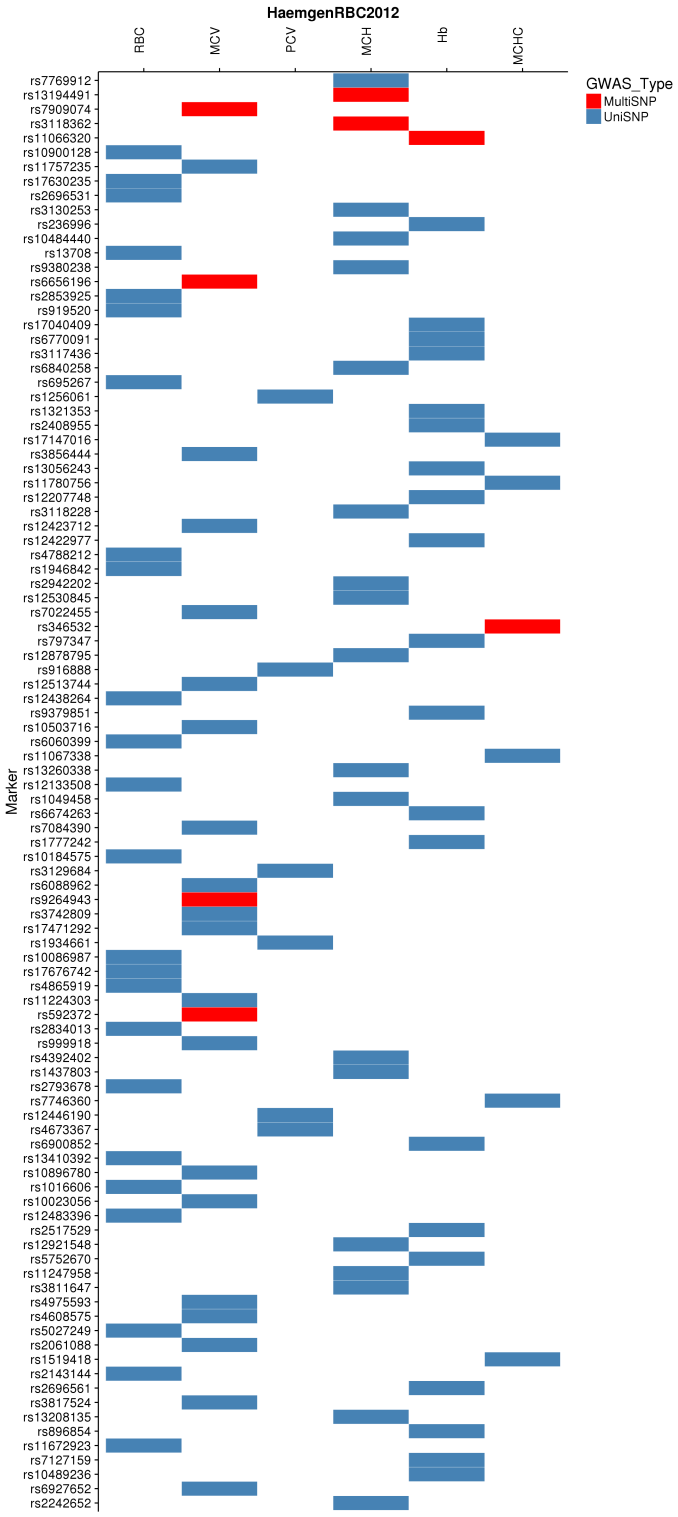

Figure 4.18: **HaemgenRBC2012 NewSNPs Significance Ranks**

Figure 4.18 (Cont.): **HaemgenRBC2012 NewSNPs Significance Ranks** – Distribution of NewSNPs among the ranking of marginally significant SNPs for HaemgenRBC2012. See GlobalLipids2013 (4.2) for full description.

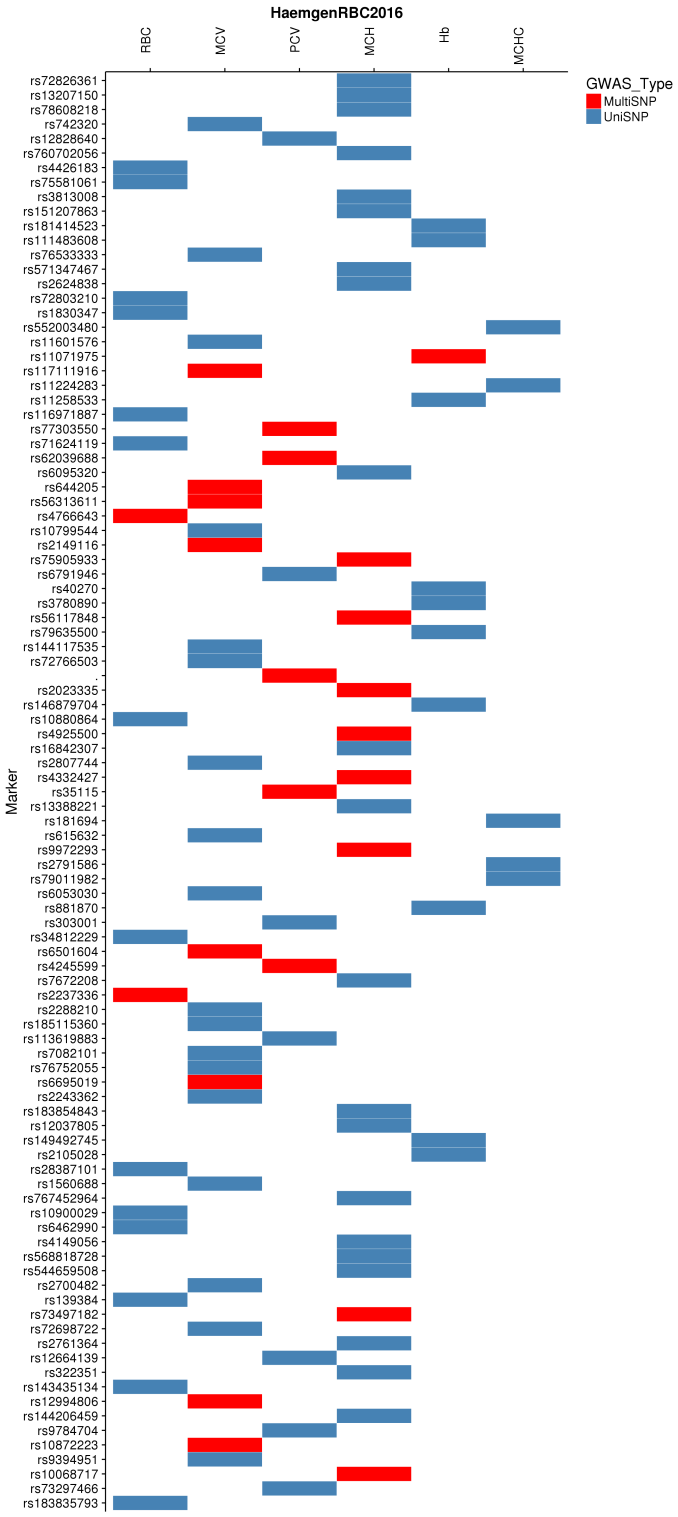

Figure 4.19: **HaemgenRBC2016 NewSNPs Significance Ranks**

Figure 4.19 (Cont.): **HaemgenRBC2016 NewSNPs Significance Ranks** – Distribution of NewSNPs among the ranking of marginally significant SNPs for HaemgenRBC2016. See GlobalLipids2013 (4.2) for full description.

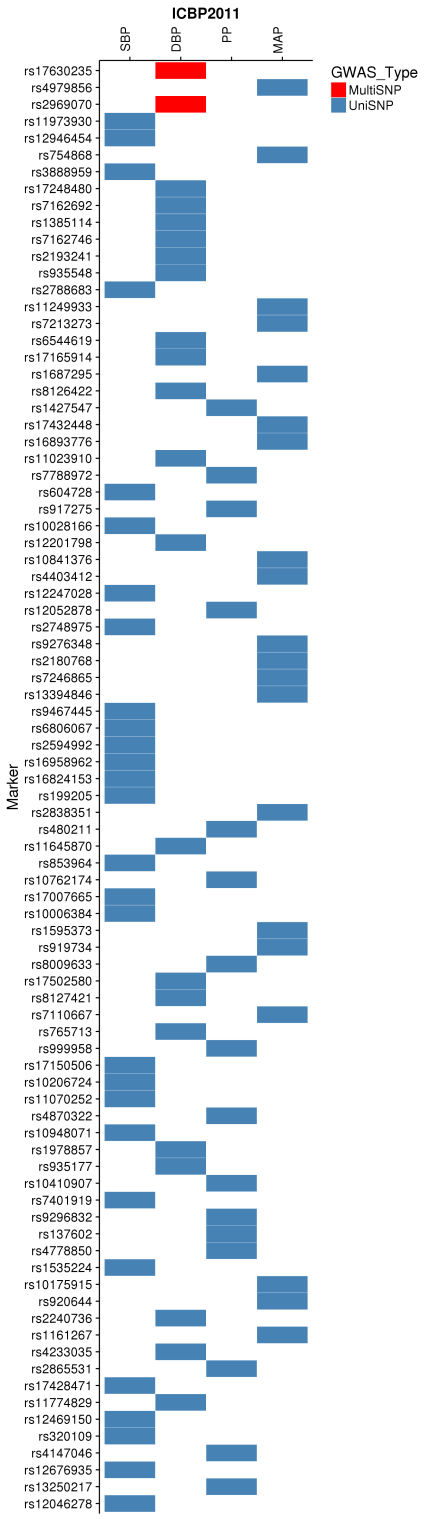

Figure 4.20: **ICBP2011 NewSNPs Significance Ranks**

Figure 4.20 (Cont.): **ICBP2011 NewSNPs Significance Ranks** – Distribution of NewSNPs among the ranking of marginally significant SNPs for ICBP2011. See GlobalLipids2013 (4.2) for full description.

MAGIC2010 NewSNPs Significance Ranks – No figure is presented for MAGIC2010 as there are no NewSNPs.

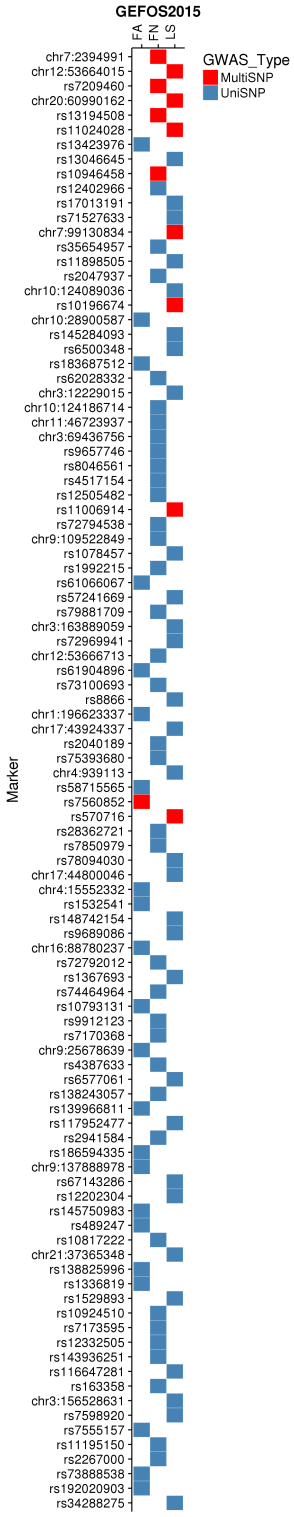

Figure 4.21: **GEFOS2015 NewSNPs Significance Ranks**

Figure 4.21 (Cont.): **GEFOS2015 NewSNPs Significance Ranks** – Distribution of NewSNPs among the ranking of marginally significant SNPs for GEFOS2015. See GlobalLipids2013 (4.2) for full description.

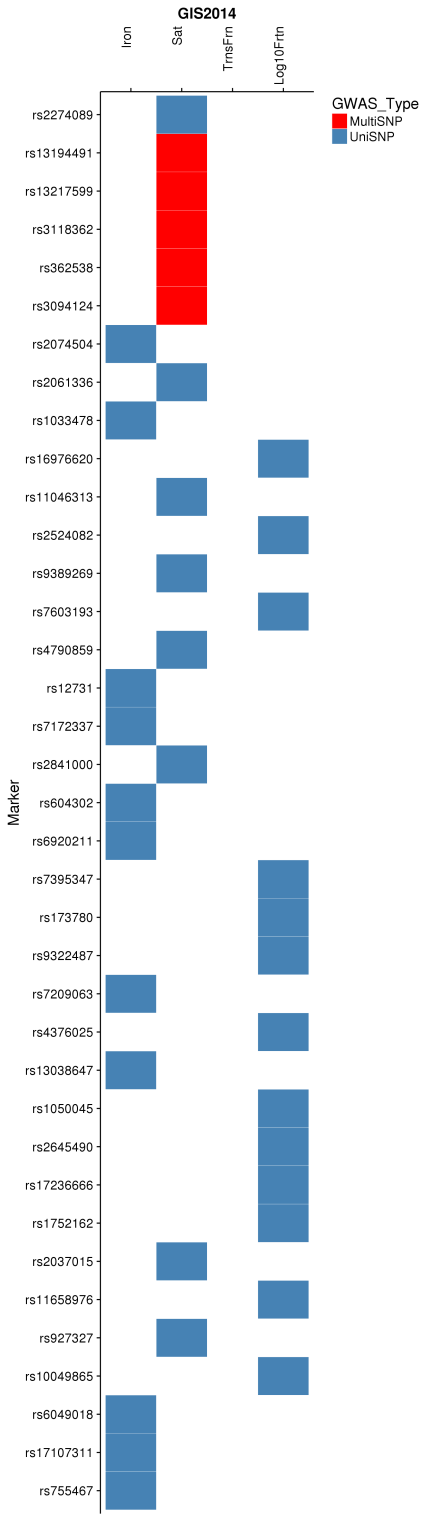

Figure 4.22: **GIS2014 NewSNPs Significance Ranks**

Figure 4.22 (Cont.): **GIS2014 NewSNPs Significance Ranks** – Distribution of NewSNPs among the ranking of marginally significant SNPs for GIS2014 NewSNPs and PreviousSNPs. See GlobalLipids2013 (4.2) for full description.

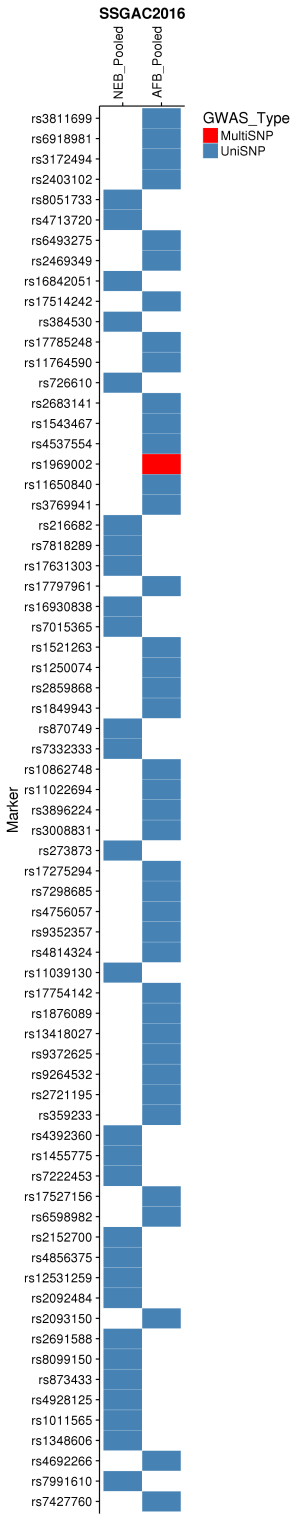

Figure 4.23: **SSGAC2016 NewSNPs Significance Ranks**

Figure 4.23 (Cont.): **SSGAC2016 NewSNPs Significance Ranks** – Distribution of NewSNPs among the ranking of marginally significant SNPs for SSGAC2016. See GlobalLipids2013 (4.2) for full description.

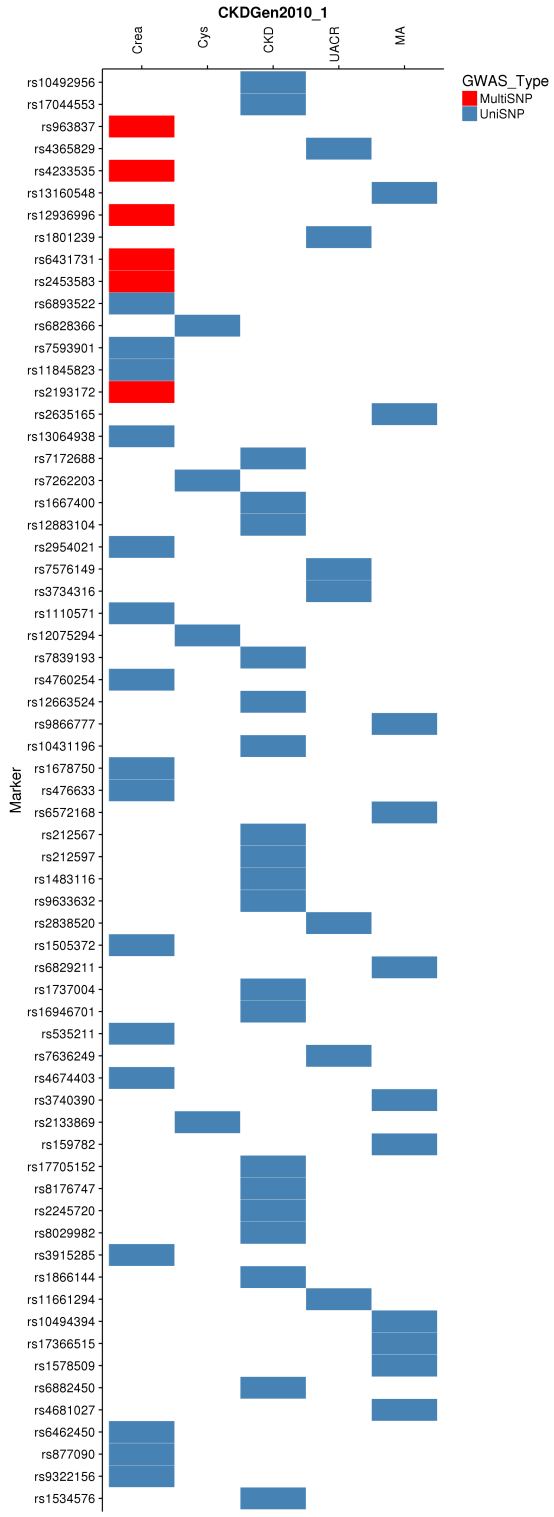

Figure 4.24: **CKDGen2010_1 NewSNPs Significance Ranks**

Figure 4.24 (Cont.): **CKDGen2010_1 NewSNPs Significance Ranks** – Distribution of NewSNPs among the ranking of marginally significant SNPs for CKDGen2010_1. See GlobalLipids2013 (4.2) for full description.

EMERGE22015 NewSNPs Significance Ranks – No figure is presented for EMERGE22015 as there is only a single NewSNP.

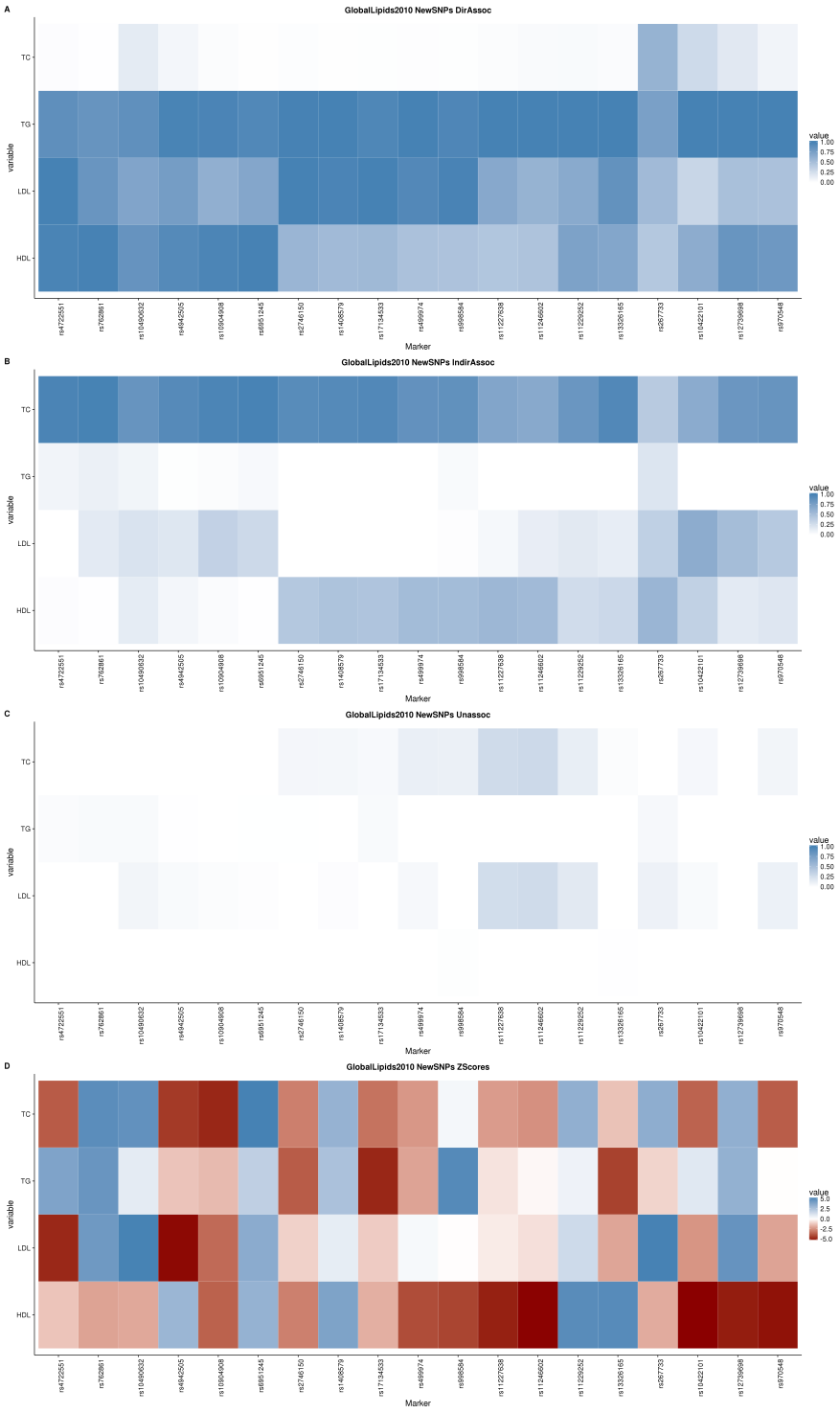

Figure 4.25: **GlobalLipids2010 NewSNPs Marginal Posteriors**

Figure 4.25 (Cont.): **GlobalLipids2010 NewSNPs Marginal Posteriors** – Marginal posterior probabilities of each NewSNP-phenotype combination being classified as **D**, **I**, or **U** for GlobalLipids2010. See GIANT2014/5 (4.3) for full description.

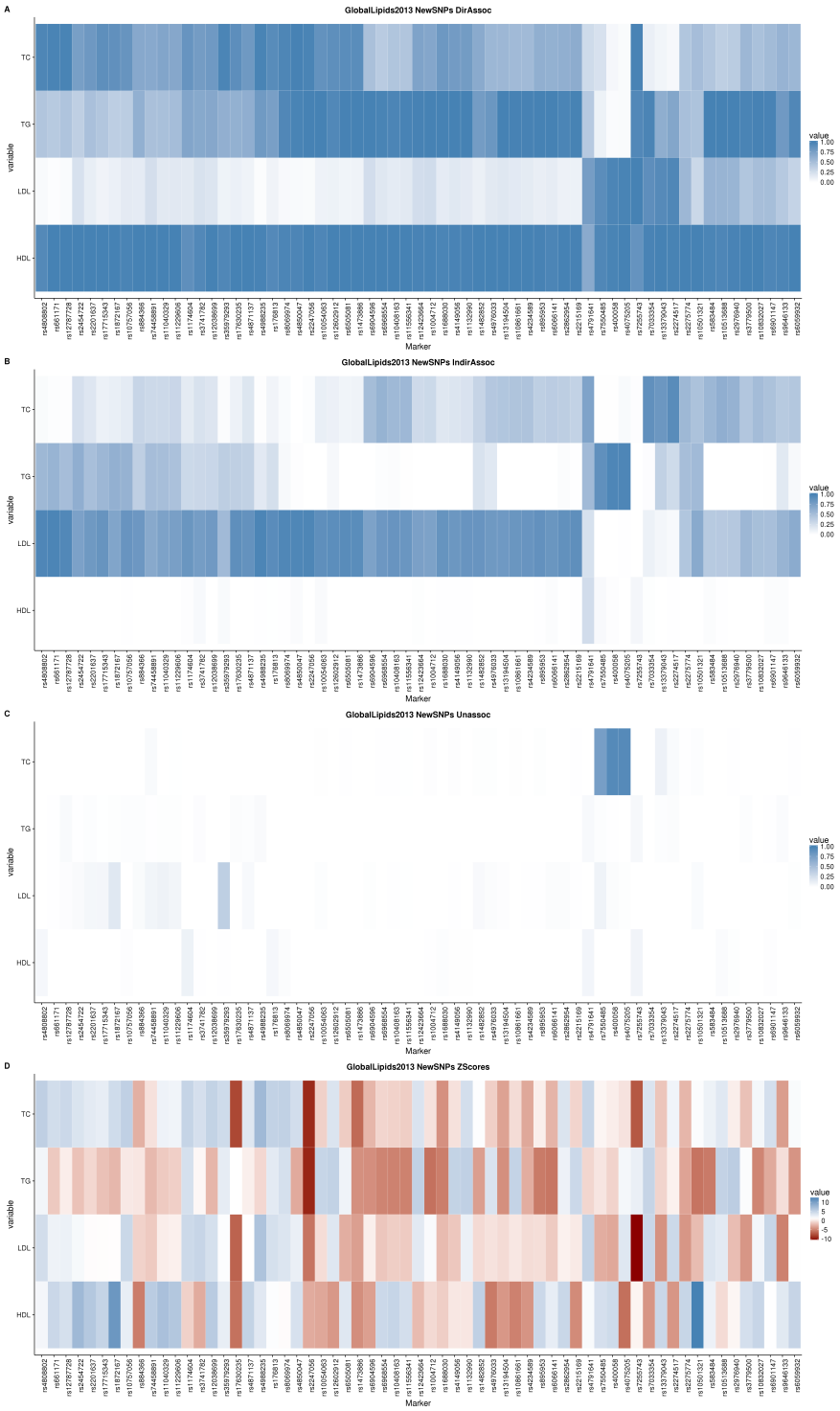

Figure 4.26: **GlobalLipids2013 NewSNPs Marginal Posteriors**

Figure 4.26 (Cont.): **GlobalLipids2013 NewSNPs Marginal Posteriors** – Marginal posterior probabilities of each NewSNP-phenotype combination being classified as **D**, **I**, or **U** for GlobalLipids2013. See GIANT2014/5 (4.3) for full description.

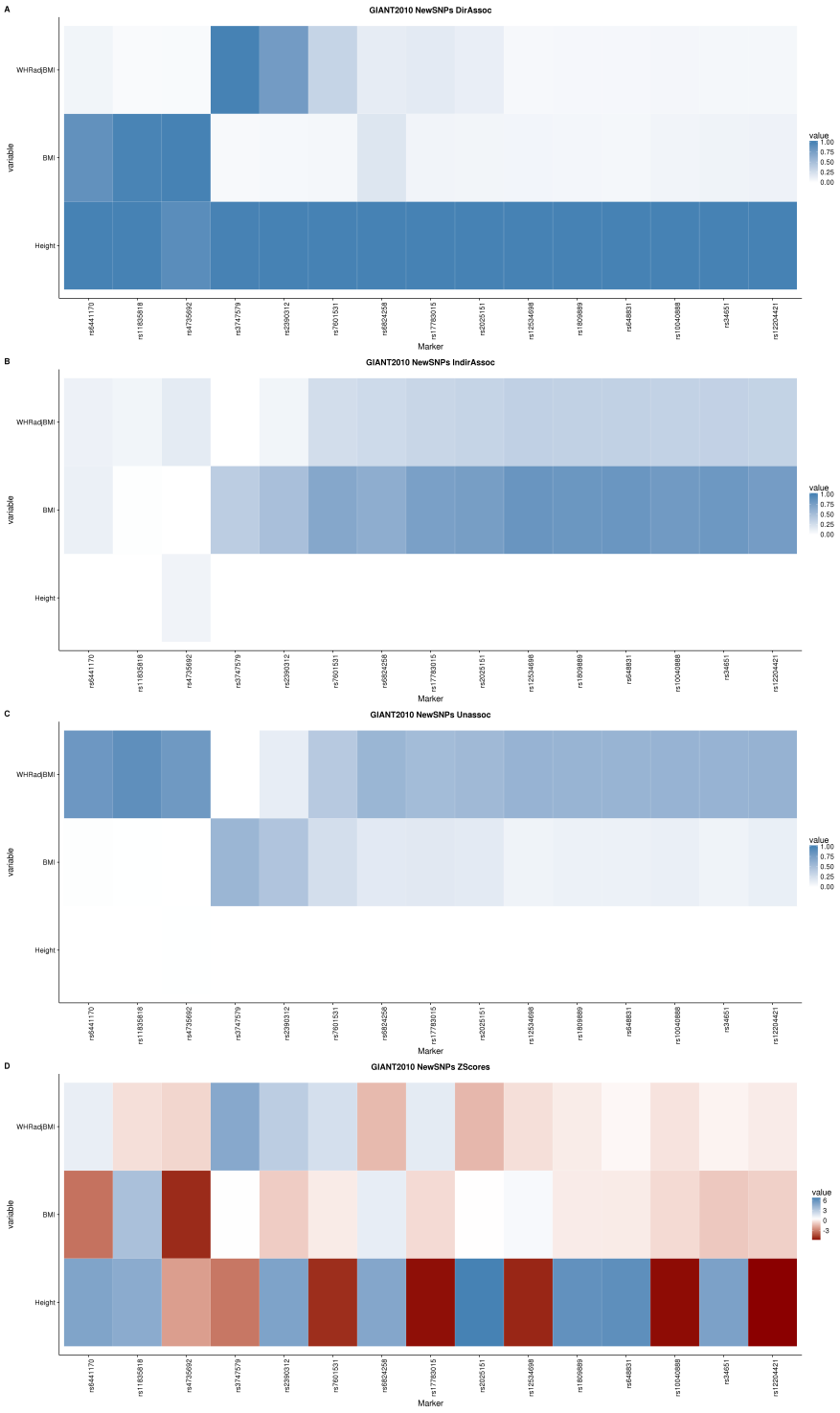

Figure 4.27: **GIANT2010 NewSNPs Marginal Posteriors**

Figure 4.27 (Cont.): **GIANT2010 NewSNPs Marginal Posteriors** – Marginal posterior probabilities of each NewSNP-phenotype combination being classified as **D**, **I**, or **U** for GIANT2010. See GIANT2014/5 (4.3) for full description.

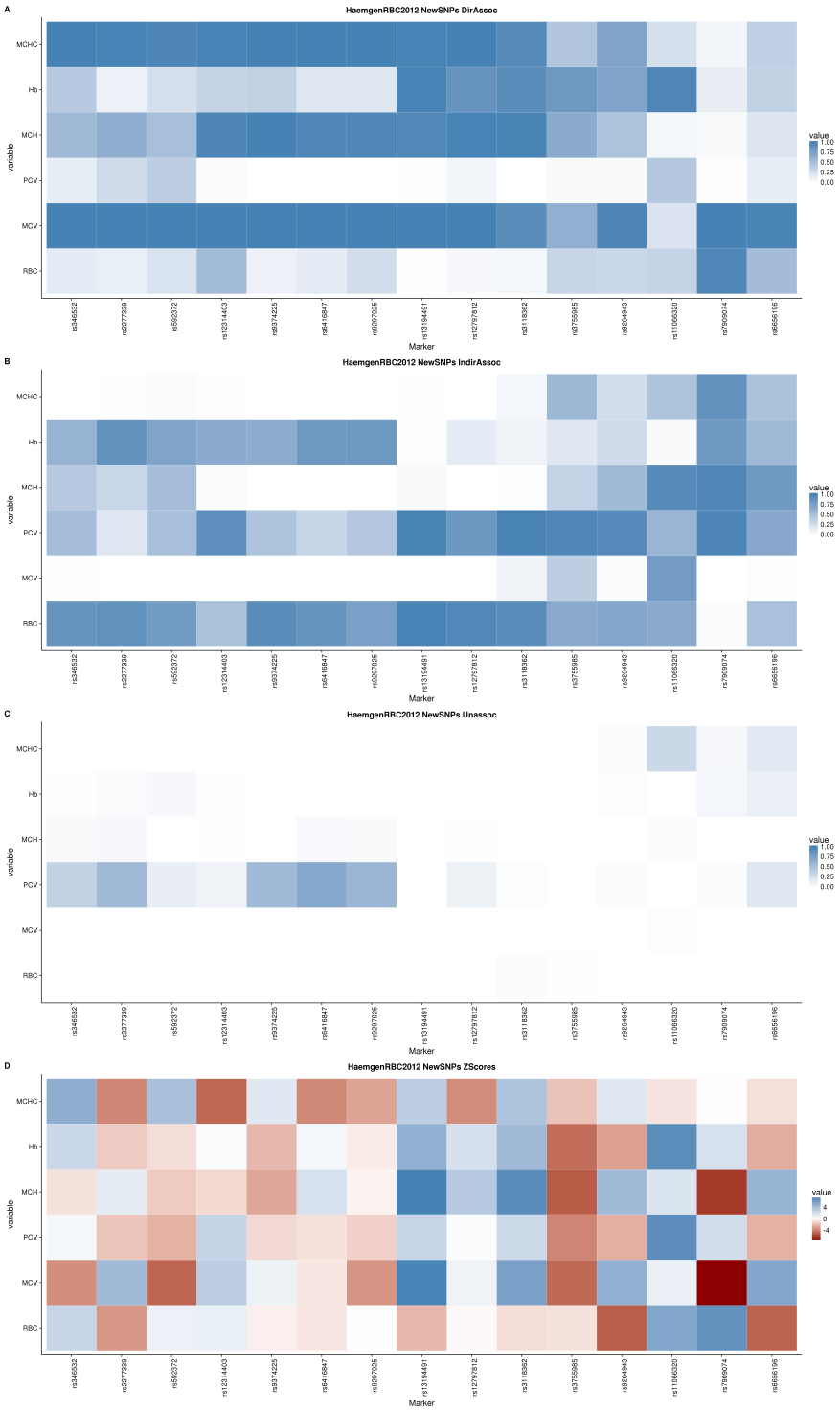

Figure 4.28: **HaemgenRBC2012 NewSNPs Marginal Posteriors**

Figure 4.28 (Cont.): **HaemgenRBC2012 NewSNPs Marginal Posteriors** – Marginal posterior probabilities of each NewSNP-phenotype combination being classified as **D**, **I**, or **U** for HaemgenRBC2012. See GIANT2014/5 (4.3) for full description.

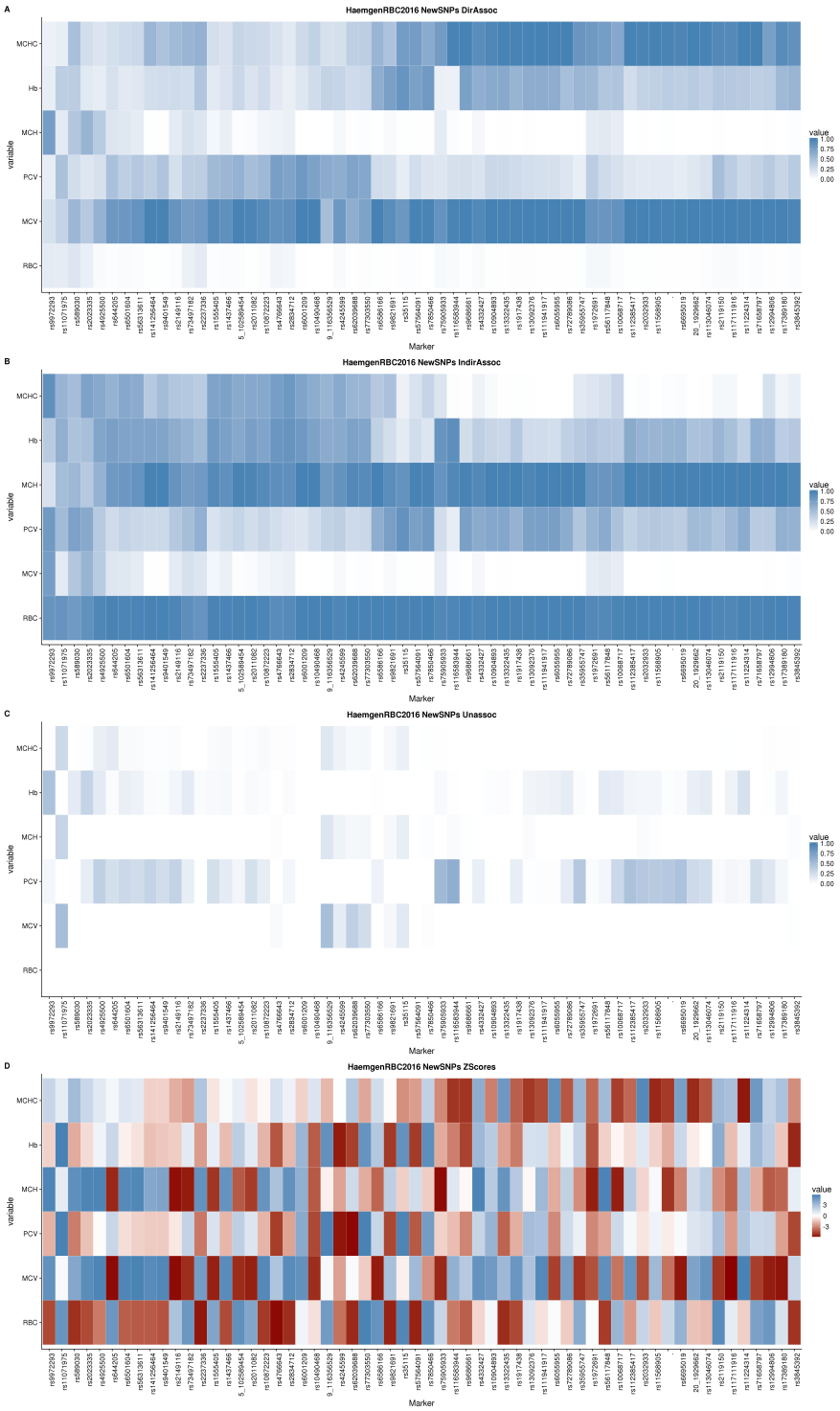

Figure 4.29: **HaemgenRBC2016 NewSNPs Marginal Posteriors**

Figure 4.29 (Cont.): **HaemgenRBC2016 NewSNPs Marginal Posteriors** – Marginal posterior probabilities of each NewSNP-phenotype combination being classified as **D**, **I**, or **U** for HaemgenRBC2016. See GIANT2014/5 (4.3) for full description.

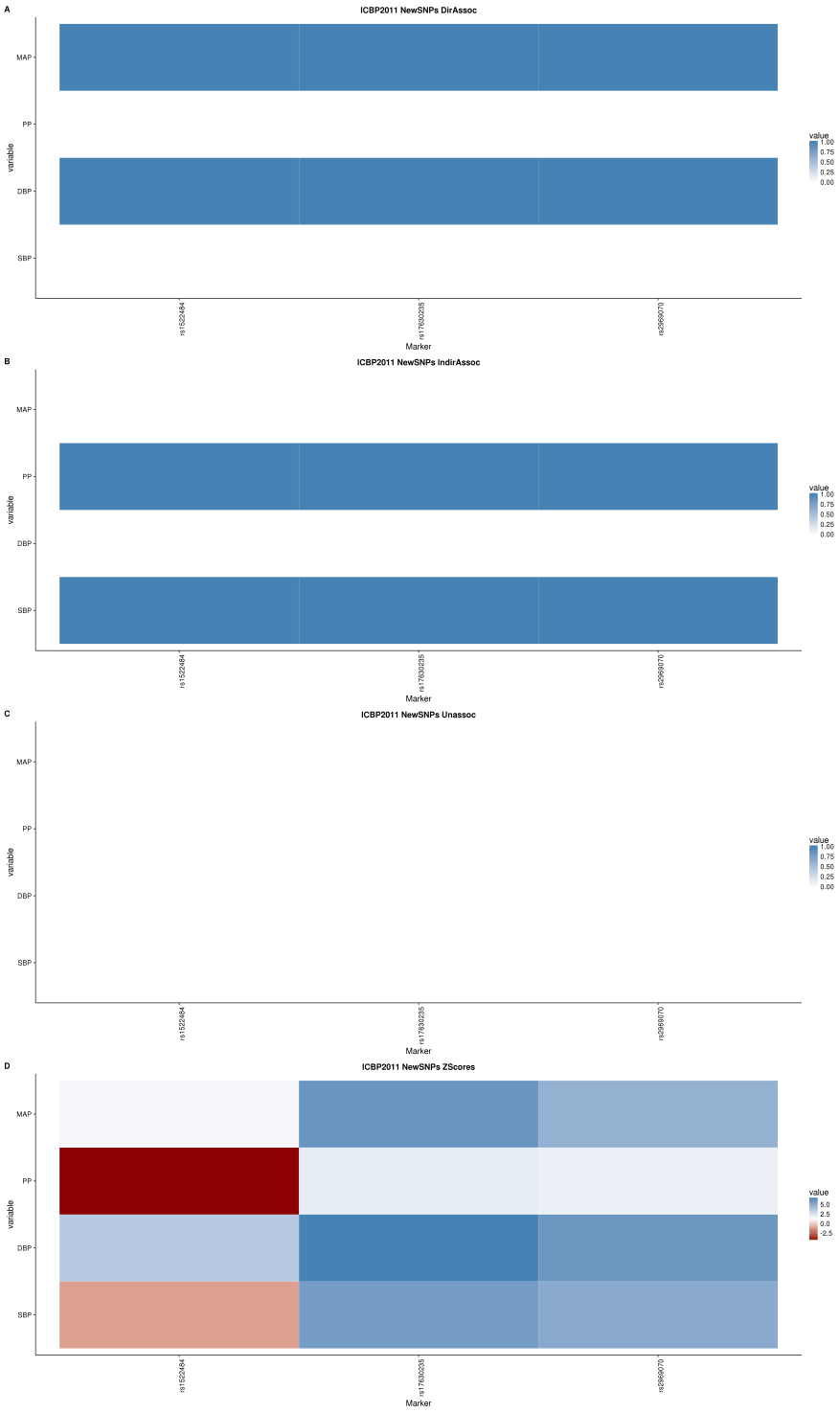

Figure 4.30: **ICBP2011 NewSNPs Marginal Posteriors**

Figure 4.30 (Cont.): **ICBP2011 NewSNPs Marginal Posteriors** – Marginal posterior probabilities of each NewSNP-phenotype combination being classified as **D**, **I**, or **U** for ICBP2011. See GIANT2014/5 (4.3) for full description.

MAGIC2010 NewSNPs Marginal Posteriors – No figure is presented for MAGIC2010 as there are no NewSNPs.

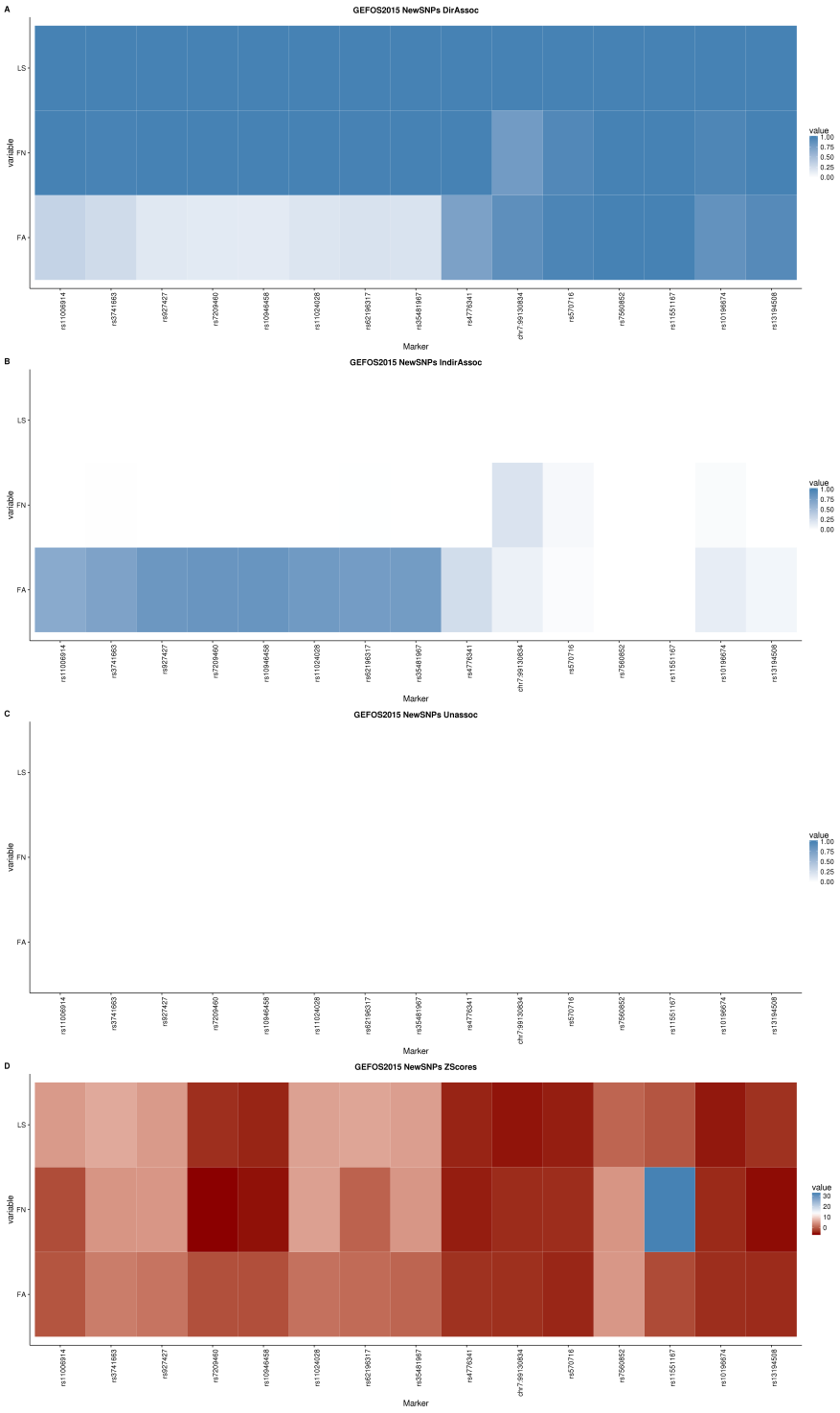

Figure 4.31: **GEFOS2015 NewSNPs Marginal Posteriors**

Figure 4.31 (Cont.): **GEFOS2015 NewSNPs Marginal Posteriors** – Marginal posterior probabilities of each NewSNP-phenotype combination being classified as **D**, **I**, or **U** for GEFOS2015. See GIANT2014/5 (4.3) for full description.

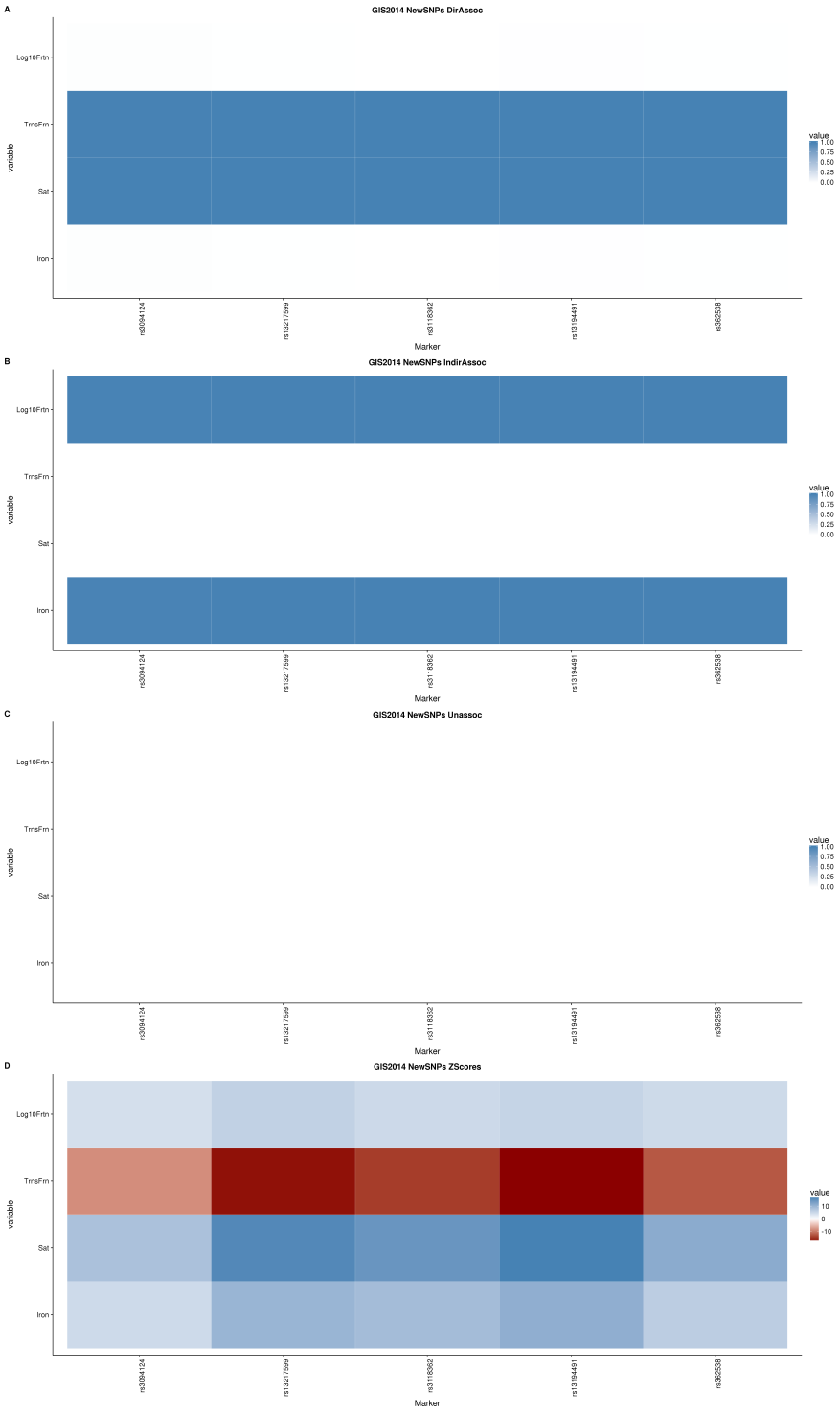

Figure 4.32: **GIS2014 NewSNPs Marginal Posteriors**

Figure 4.32 (Cont.): **GIS2014 NewSNPs Marginal Posteriors** – Marginal posterior probabilities of each NewSNP-phenotype combination being classified as **D**, **I**, or **U** for GIS2014. See GIANT2014/5 (4.3) for full description.

SSGAC2016 NewSNPs Marginal Posteriors – No figure is presented for SSGAC2016 as there is only a single NewSNP.

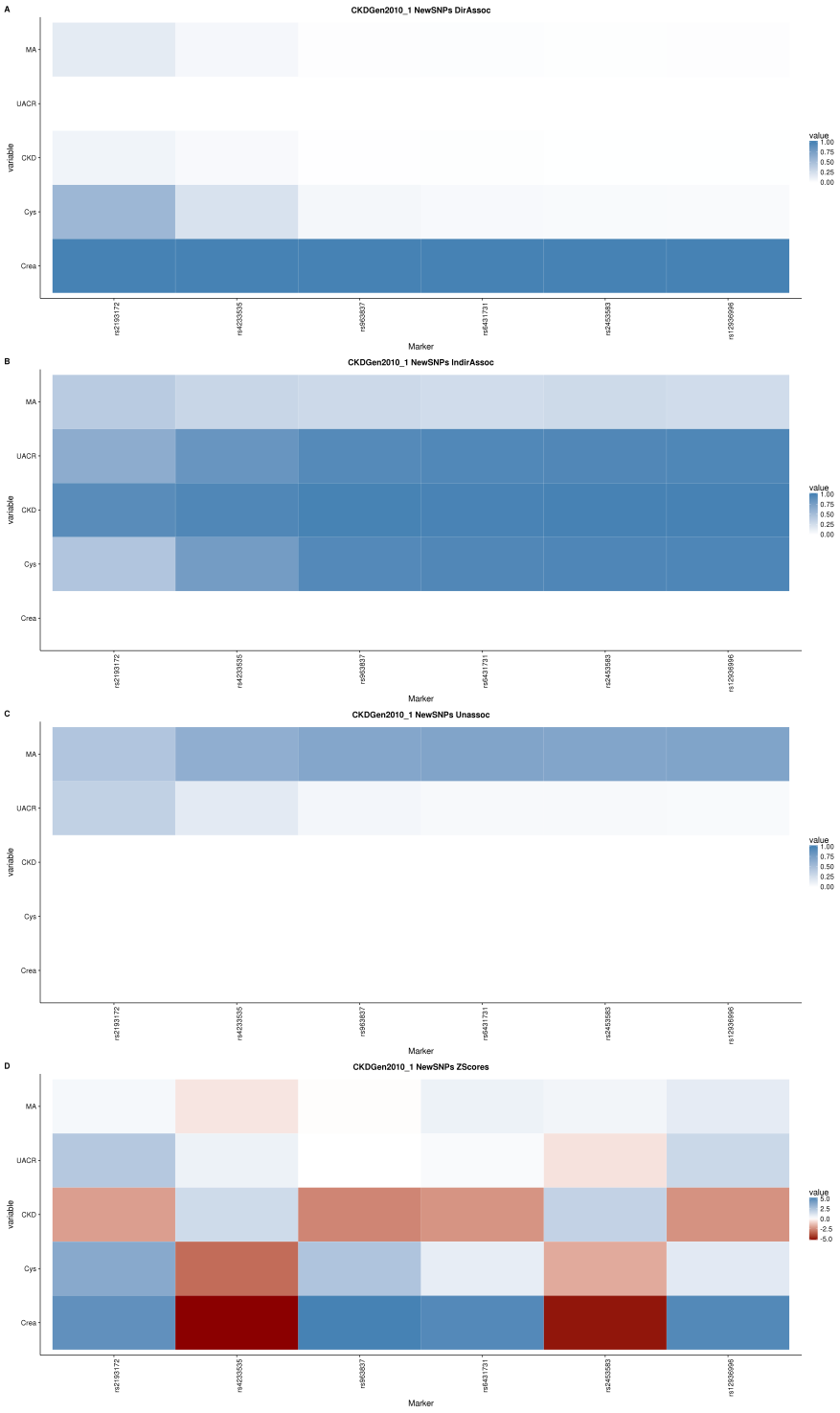

Figure 4.33: **CKDGen2010_1 NewSNPs Marginal Posteriors**

Figure 4.33 (Cont.): **CKDGen2010_1 NewSNPs Marginal Posteriors** – Marginal posterior probabilities of each NewSNP-phenotype combination being classified as **D**, **I**, or **U** for CKDGen2010_1. See GIANT2014/5 (4.3) for full description.

EMERGE22015 NewSNPs Marginal Posteriors – No figure is presented for SSGAC2016 as there is only a single NewSNP.

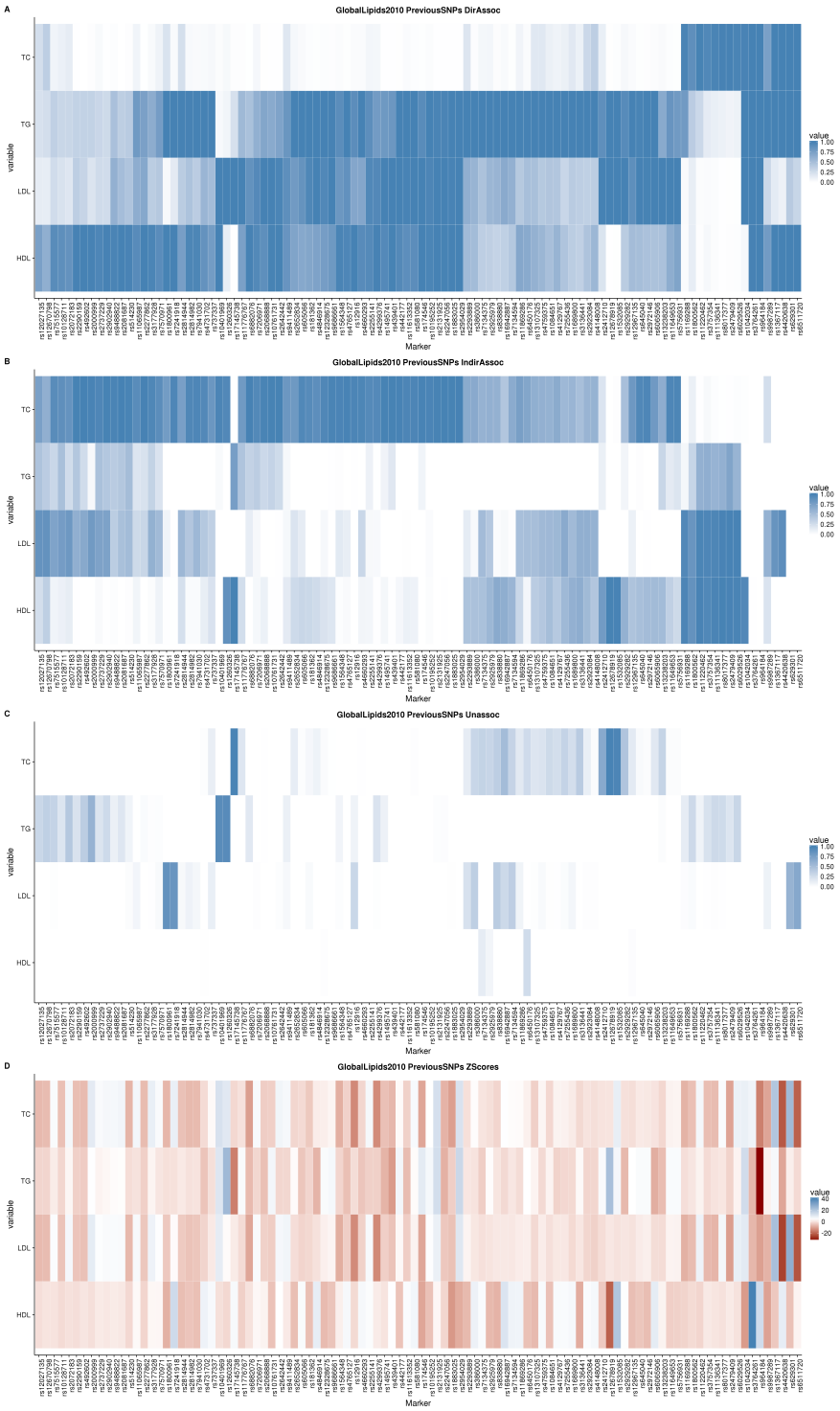

Figure 4.34: **GlobalLipids2010 PreviousSNPs Marginal Posteriors**

Figure 4.34 (Cont.): **GlobalLipids2010 PreviousSNPs Marginal Posteriors** – Marginal posterior probabilities of each PreviousSNP-phenotype combination being classified as **D**, **I**, or **U** for GlobalLipids2010. See GIANT2014/5 4.3 for full description.

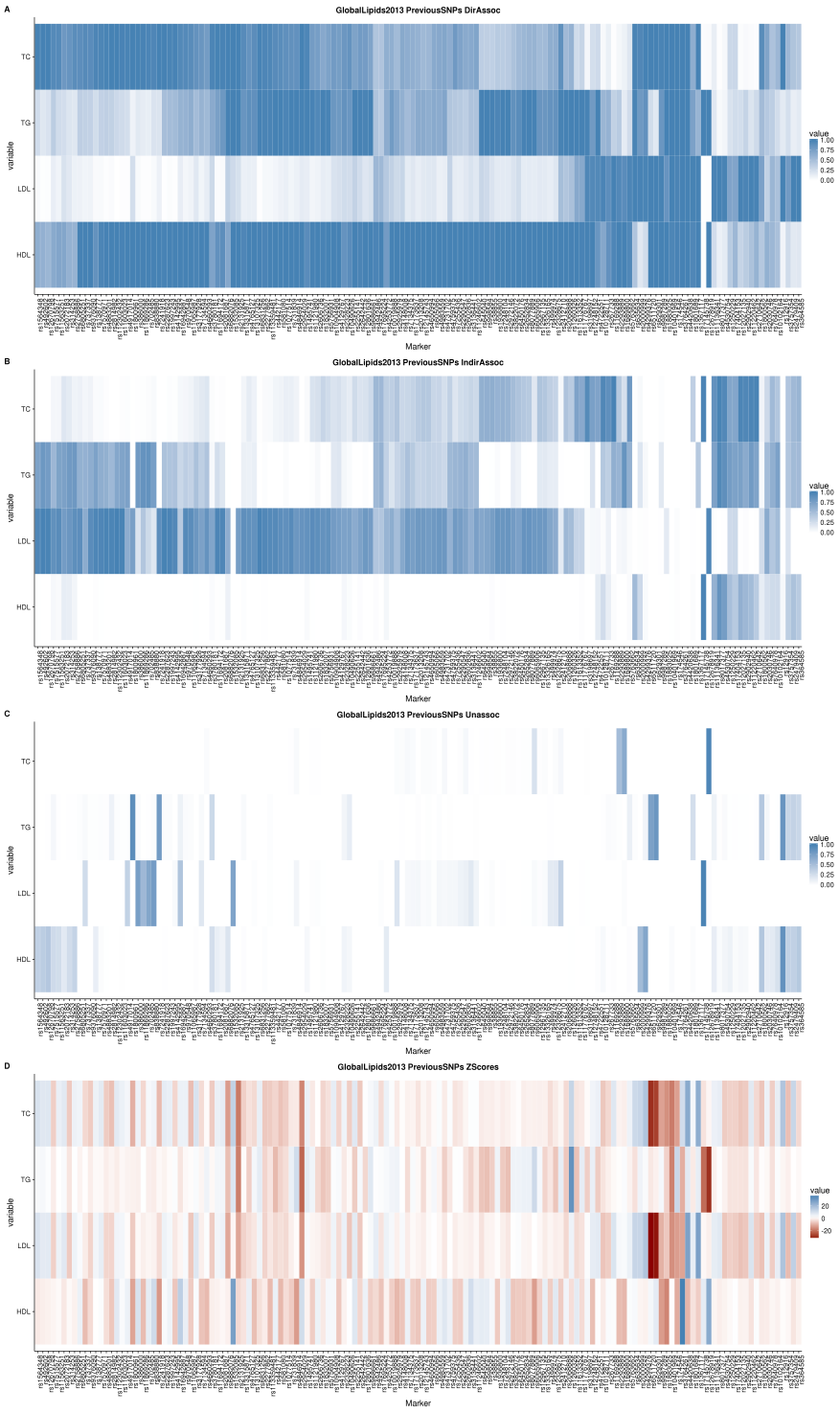

Figure 4.35: **GlobalLipids2013 PreviousSNPs Marginal Posteriors**

Figure 4.35 (Cont.): **GlobalLipids2013 PreviousSNPs Marginal Posteriors** – Marginal posterior probabilities of each PreviousSNP-phenotype combination being classified as **D**, **I**, or **U** for GlobalLipids2013. See GIANT2014/5 (4.3) for full description.

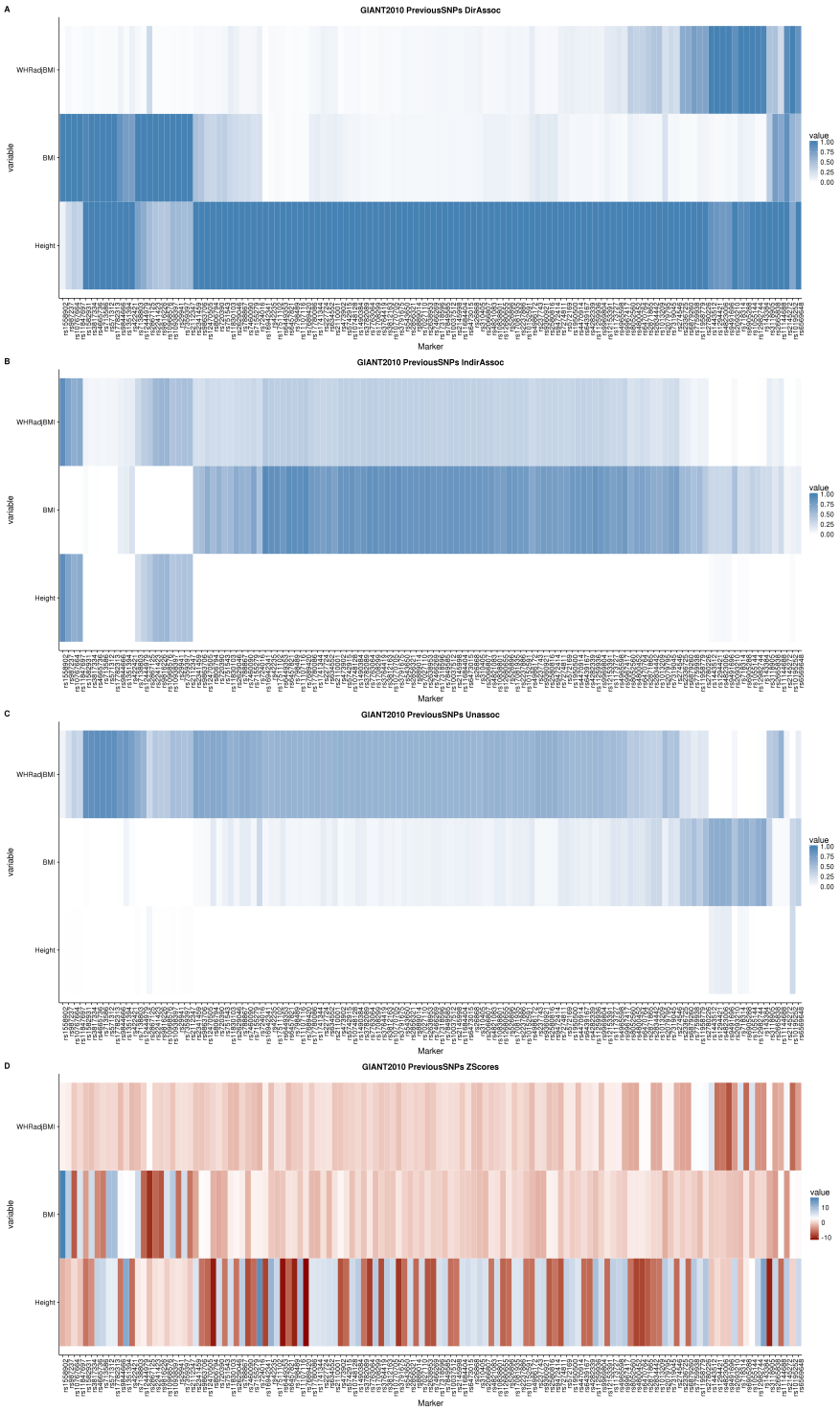

Figure 4.36: **GIANT2010 PreviousSNPs Marginal Posteriors**

Figure 4.36 (Cont.): **GIANT2010 PreviousSNPs Marginal Posteriors** – Marginal posterior probabilities of each PreviousSNP-phenotype combination being classified as **D**, **I**, or **U** for GIANT2010. See GIANT2014/5 (4.3) for full description.

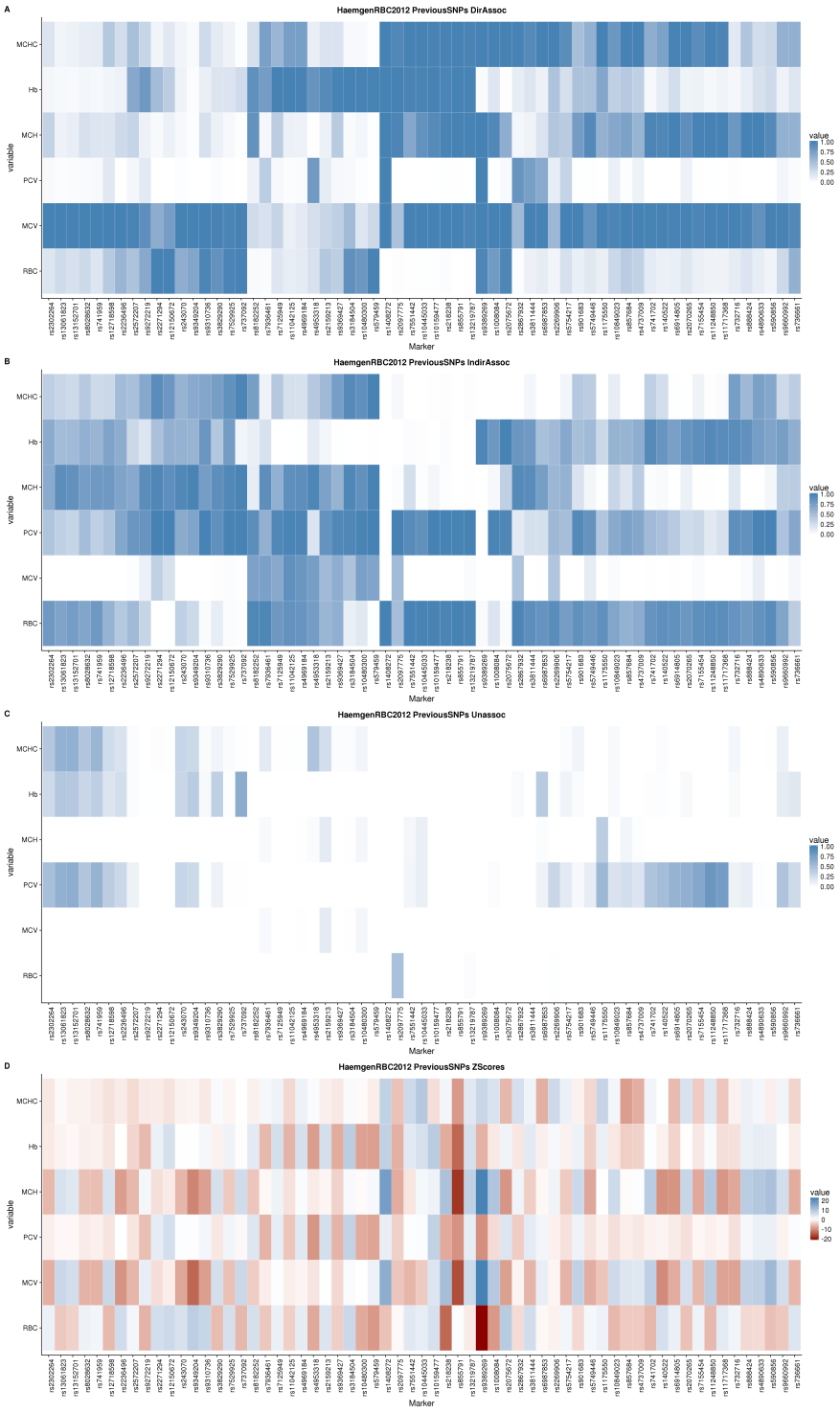

Figure 4.37: **HaemgenRBC2012 PreviousSNPs Marginal Posteriors**

Figure 4.37 (Cont.): **HaemgenRBC2012 PreviousSNPs Marginal Posteriors** – Marginal posterior probabilities of each PreviousSNP-phenotype combination being classified as **D**, **I**, or **U** for HaemenRBC2012. See GIANT2014/5 (4.3) for full description.

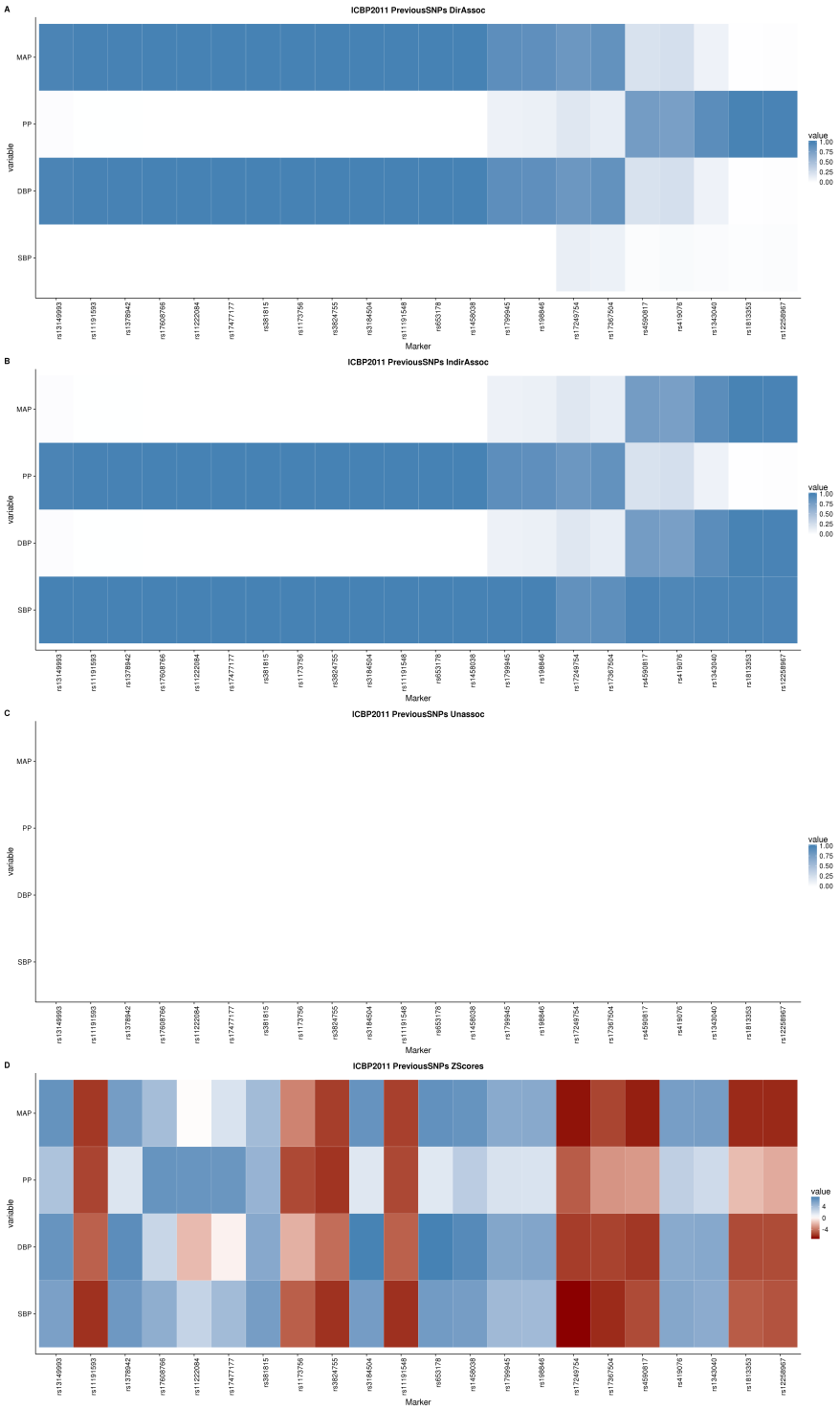

Figure 4.38: **ICBP2011 PreviousSNPs Marginal Posteriors**

Figure 4.38 (Cont.): **ICBP2011 PreviousSNPs Marginal Posteriors** – Marginal posterior probabilities of each PreviousSNP-phenotype combination being classified as **D**, **I**, or **U** for ICBP2011. See GIANT2014/5 (4.3) for full description.

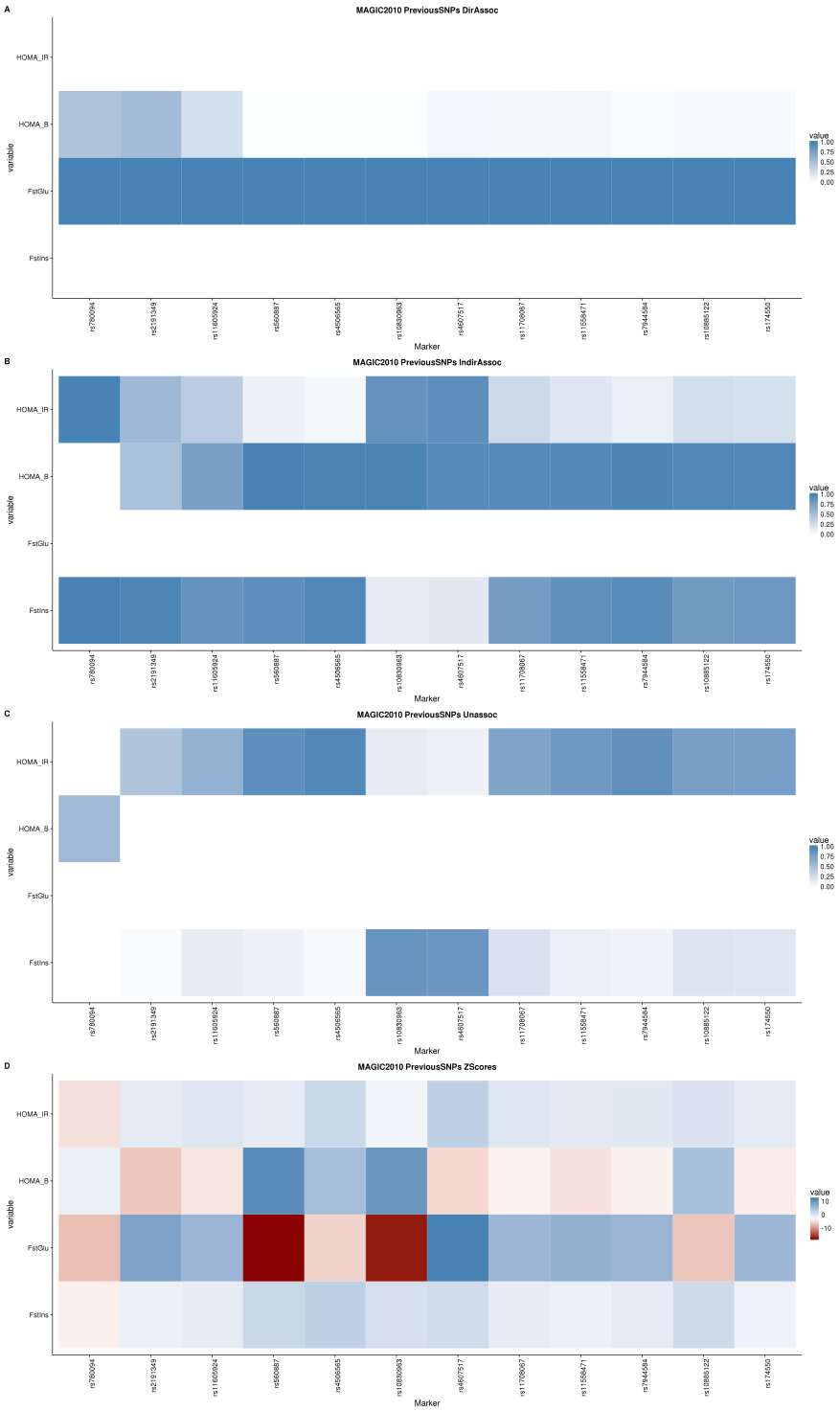

Figure 4.39: **MAGIC2010 PreviousSNPs Marginal Posteriors**

Figure 4.39 (Cont.): **MAGIC2010 PreviousSNPs Marginal Posteriors** – Marginal posterior probabilities of each PreviousSNP-phenotype combination being classified as **D**, **I**, or **U** for MAGIC2010. See GIANT2014/5 (4.3) for full description.

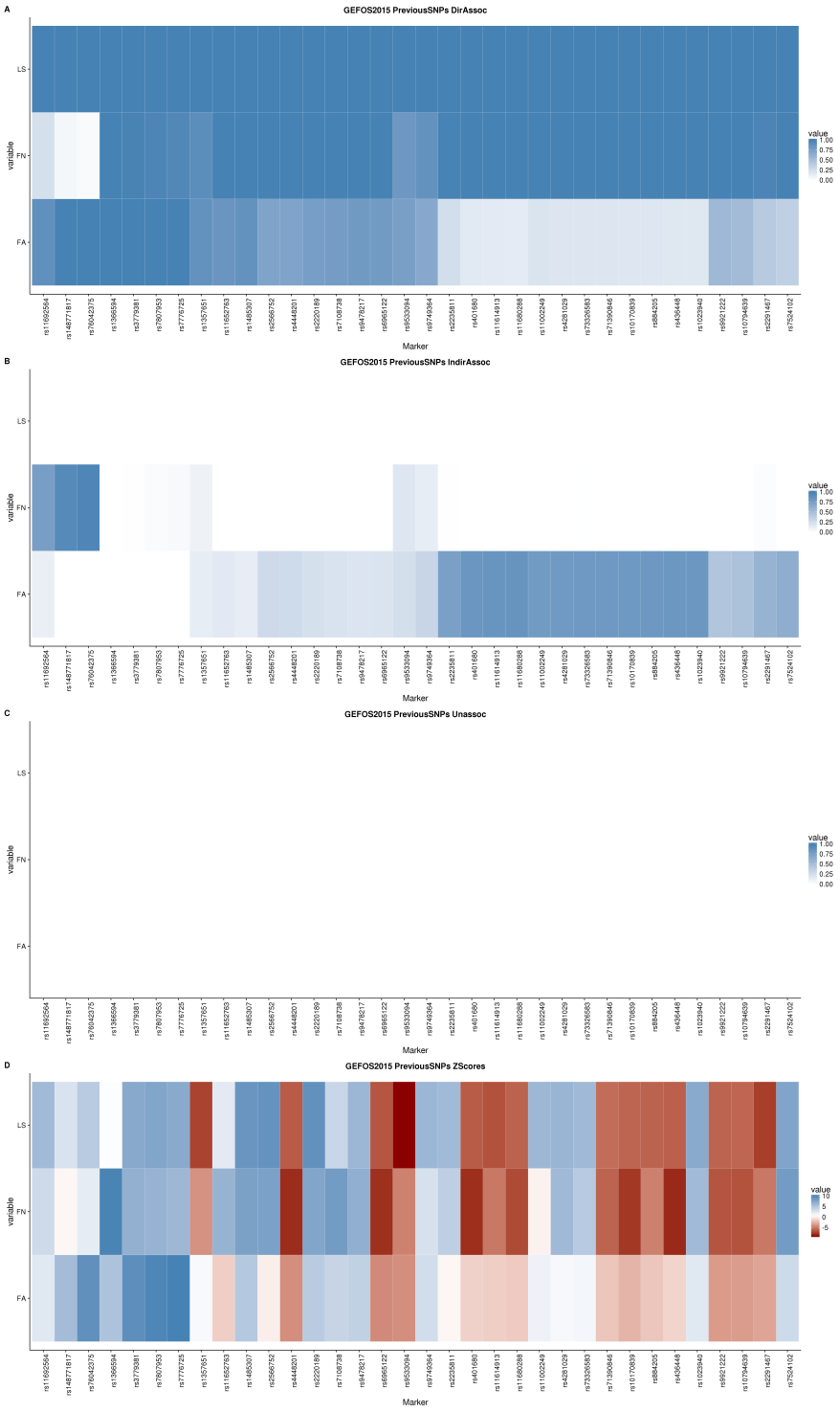

Figure 4.40: **GEFOS2015 PreviousSNPs Marginal Posteriors**

Figure 4.40 (Cont.): **GEFOS2015 PreviousSNPs Marginal Posteriors** – Marginal posterior probabilities of each PreviousSNP-phenotype combination being classified as **D**, **I**, or **U** for GEFOS2015. See GIANT2014/5 (4.3) for full description.

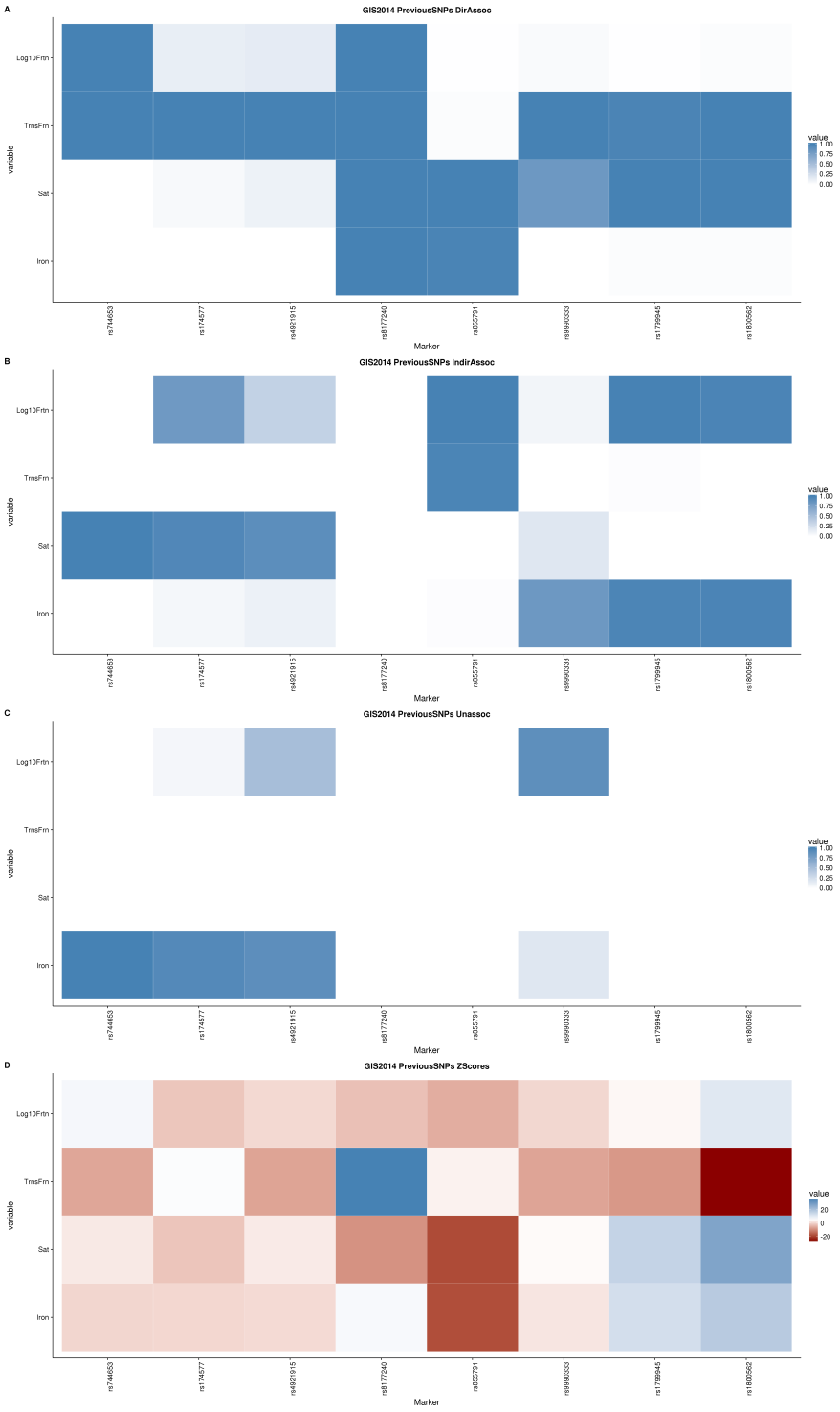

Figure 4.41: **GIS2014 PreviousSNPs Marginal Posteriors**

Figure 4.41 (Cont.): **GIS2014 PreviousSNPs Marginal Posteriors** – Marginal posterior probabilities of each PreviousSNP-phenotype combination being classified as **D**, **I**, or **U** for GIS2014. See GIANT2014/5 (4.3) for full description.

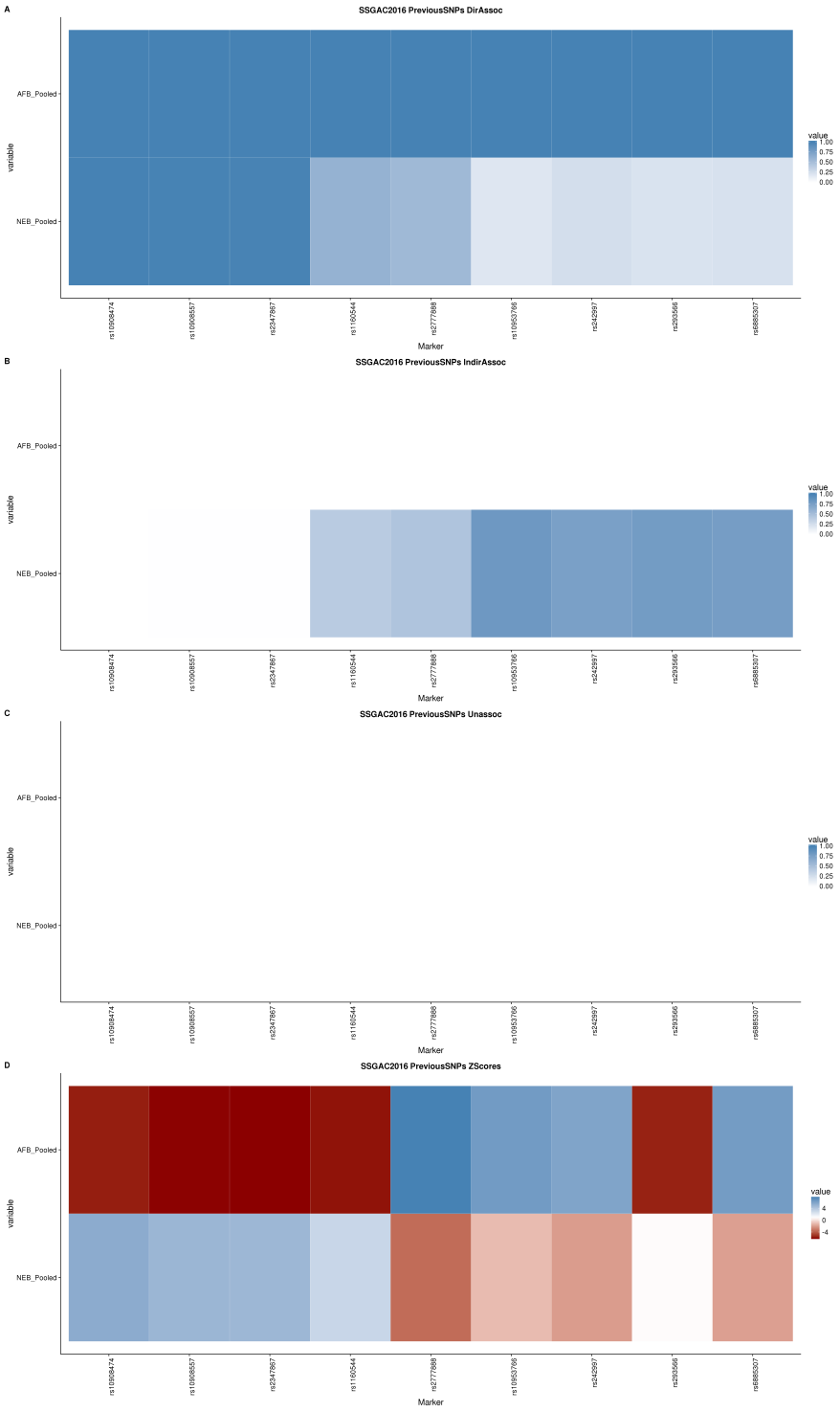

Figure 4.42: **SSGAC2016 PreviousSNPs Marginal Posteriors**

Figure 4.42 (Cont.): **SSGAC2016 PreviousSNPs Marginal Posteriors** – Marginal posterior probabilities of each PreviousSNP-phenotype combination being classified as **D**, **I**, or **U** for SSGAC2016. See GIANT2014/5 (4.3) for full description.

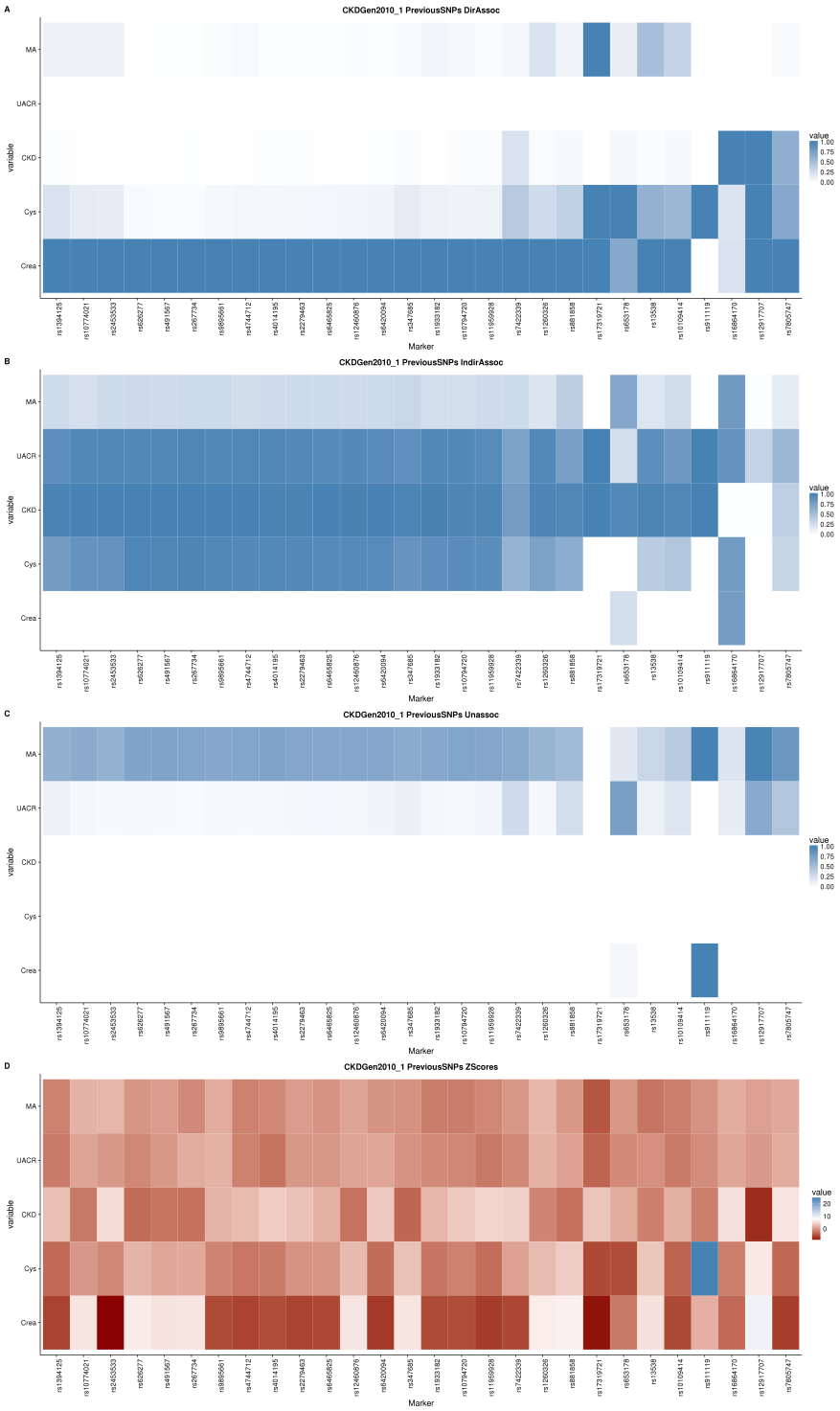

Figure 4.43: **CKDGen2010_1 PreviousSNPs Marginal Posteriors**

Figure 4.43 (Cont.): **CKDGen2010_1 PreviousSNPs Marginal Posteriors** – Marginal posterior probabilities of each PreviousSNP-phenotype combination being classified as **D**, **I**, or **U** for CKDGen2010_1. See GIANT2014/5 (4.3) for full description.

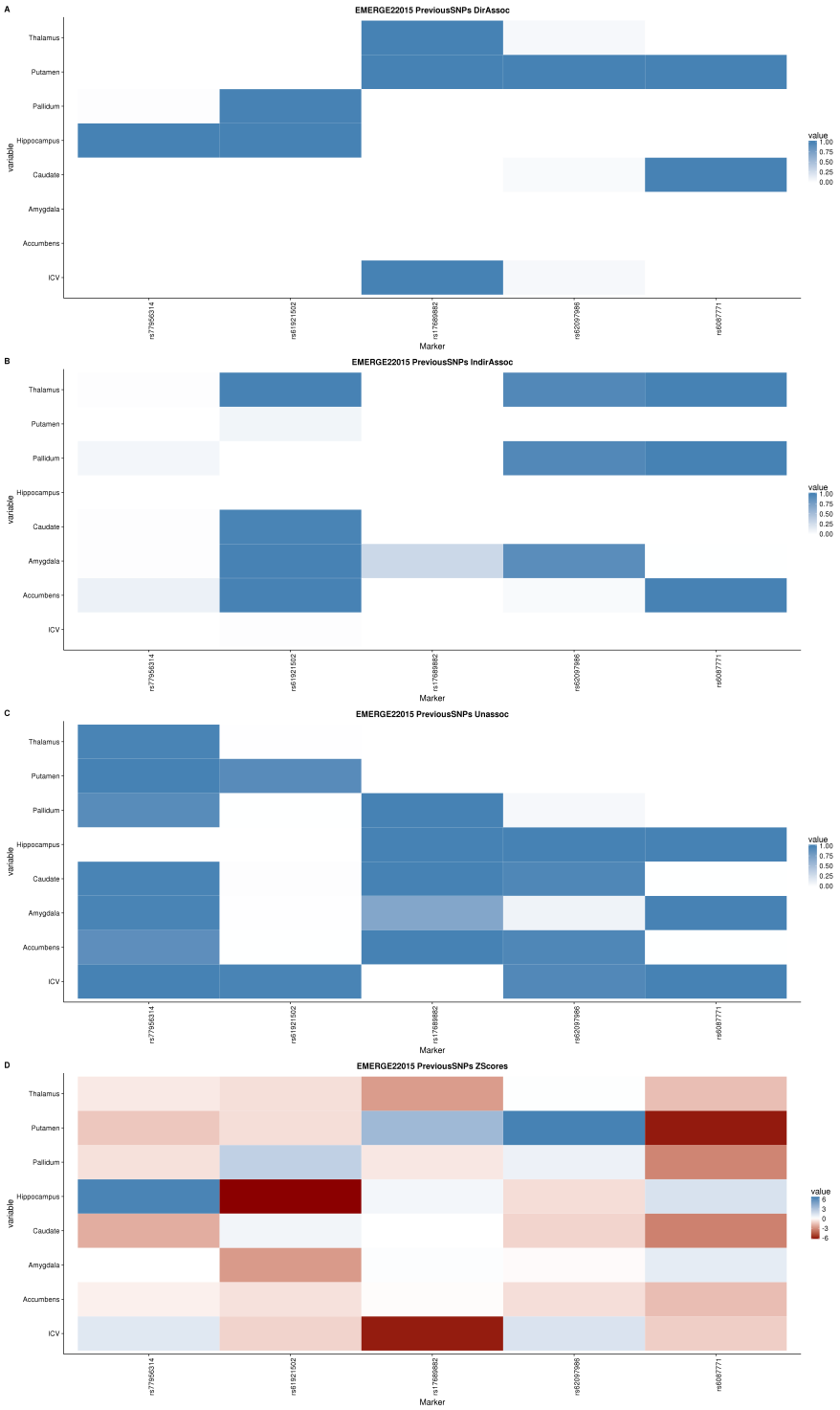

Figure 4.44: **EMERGE22015 PreviousSNPs Marginal Posteriors**

Figure 4.44 (Cont.): **EMERGE22015 PreviousSNPs Marginal Posteriors** – Marginal posterior probabilities of each PreviousSNP-phenotype combination being classified as **D**, **I**, or **U** for EMERGE22015. See GIANT2014/5 (4.3) for full description.

4.11 Supplementary Tables

Dataset	Release	Previous SNPs Total	Previous SNPs Used	PreviousSNPs Dropped	PreviousSNPs GWASThresh
GlobalLipids	2010	102	100	2	0
	2013	157	145	12	2
GIANT	2010	226	128	98	46
	2014/5	843	724	119	66
HaemgenRBC	2012	75	63	12	12
	2016	642	610	32	0
ICBP	2011	49	22	27	27
MAGIC	2010	17	12	5	3
GEFOS	2015	36	35	1	0
GIS	2014	12	8	4	4
SSGAC	2016	11	9	2	2
CKDGen	2010/1	28	28	0	0
EMERGE2	2015	8	5	3	2

Table 4.7: **Summary of Datasets' PreviousSNPs and related metrics** – List showing for each dataset different metrics regarding that study's PreviousSNPs. First column shows the original number of PreviousSNPs retrieved from across all phenotypes in the study (see Online Methods for sources per dataset). Second column shows the number of PreviousSNPs that made it to the final joined dataset that was used for analysis. Third column shows the number of PreviousSNPs that were dropped in the process of creating the final joined files; SNPs were droped for a variety of reasons, including not having a phenotype measured in all phenotypes analyzed, having a MAF equal to zero, having non-matching reference alleles across phenotypes, and not passing the study's original univariate GWAS p-value significance threshold. The final column shows the number of PreviousSNPs that were dropped for this last mentioned reason, not passing the univariate GWAS p-value threshold. This indicates there were PreviousSNPs originally reported by the study that were identified by the inclusion of data beyond the publicly relased discovery set; this column is also displayed in Main Table 4.3.

Chr	BP	Marker	MAF	A1	unistat_log10pVal	mvstat_log10pVal	logBFWeightedAvg
1	27102620	rs12739698	0.08	G	6.46	8.53	7.20
1	149225460	rs267733	0.14	G	7.19	7.09	4.74
2	118295555	rs10490632	0.08	G	7.14	6.48	5.13
3	52507158	rs13326165	0.21	G	6.25	6.15	4.74
4	3411809	rs762861	0.26	G	6.04	6.73	5.24
6	29550680	rs2746150	0.09	C	3.78	6.47	5.28
6	43865874	rs998584	0.49	C	6.39	6.16	4.46
7	1024719	rs6951245	0.16	G	7.22	6.79	5.41
7	25958351	rs4722551	0.18	C	6.00	9.38	7.52
10	5237098	rs17134533	0.15	G	6.06	7.29	6.04
10	17300296	rs10904908	0.43	G	6.06	6.26	4.62
10	45333283	rs970548	0.25	C	6.81	6.77	5.44
10	101902184	rs1408579	0.47	C	4.07	6.52	4.83
11	51368666	rs11246602	0.13	C	7.16	6.34	5.31
11	54886216	rs11229252	0.09	C	6.17	6.10	4.95
11	55776161	rs11227638	0.12	T	6.30	5.83	4.79
11	75132669	rs499974	0.17	C	4.40	6.15	4.46
13	31861707	rs4942505	0.48	C	7.01	7.48	5.81
19	57011927	rs10422101	0.27	G	7.20	6.95	5.61

Table 4.8: **GlobalLipids2010 NewSNPs** – Summary statistics of new genome-wide significant SNPs (NewSNPs) identified from the bmass analysis of GlobalLipids2010. All phenotype ZScores are oriented towards the HDL minor allele (A1). Nmin is the minimum number of samples across all phenotypes for that variant. unistat_log10pVal & mvstat_log10pVal are the negative log10 p-values of the naive univariate and multivariate statistics used to determine the subset of SNPs brought forward for the final analysis (see Online Methods). logBFWeightedAvg is the final summary statistic from bmass, representing the weighted, average Bayes Factor across all models analyzed for each variant.

Chr	BP	Marker	MAF	A1	unistat_log10pVal	mvstat_log10pVal	logBFWeightedAvg
2	19831424	rs7601531	0.35	C	6.36	5.81	4.45
3	159289653	rs6441170	0.35	C	6.31	7.21	5.14
4	122989416	rs6824258	0.27	C	6.01	5.49	4.11
5	33374424	rs10040888	0.23	T	7.81	6.40	5.74
5	72179760	rs34651	0.10	C	6.60	5.25	4.64
6	33736840	rs12204421	0.22	G	8.11	6.77	6.05
6	81012926	rs648831	0.50	T	9.01	7.45	6.84
7	20348901	rs2390312	0.44	G	6.42	6.87	5.01
7	46374788	rs12534698	0.04	A	6.76	5.38	4.73
8	76778217	rs4735692	0.44	G	6.50	6.24	4.41
9	98201332	rs2025151	0.18	G	11.01	9.85	8.84
12	88755516	rs17783015	0.22	T	7.70	6.73	5.68
12	120979191	rs11835818	0.44	C	5.45	7.14	5.32
12	123367178	rs1809889	0.24	T	8.71	7.15	6.59
16	4385327	rs3747579	0.27	C	5.79	6.45	4.60

Table 4.9: **GIANT2010 NewSNPs** – Summary statistics of NewSNPs identified from the bmass analysis of GIANT2010. All phenotype ZScores are oriented towards the Height minor allele (A1). Descriptions otherwise the same as previously detailed in GlobalLipids2010 NewSNPs (4.8).

Chr	BP	Marker	MAF	A1	unistat_log10pVal	mvstat_log10pVal	logBFWeightedAvg
1	25508784	rs592372	0.42	G	5.98	6.97	5.56
1	26631360	rs6656196	0.21	A	7.53	7.35	6.30
4	88243762	rs3755985	0.25	G	6.70	7.29	5.92
6	17082682	rs9297025	0.19	G	3.02	6.65	5.43
6	27145059	rs13194491	0.08	T	14.60	16.80	14.70
6	28893064	rs3118362	0.09	C	12.25	11.95	10.56
6	31382500	rs9264943	0.12	A	6.62	7.05	5.91
6	111387025	rs9374225	0.08	C	2.15	6.56	5.37
10	44715845	rs7909074	0.41	G	13.45	13.46	13.23
11	73912984	rs12797812	0.03	T	3.51	8.48	6.27
12	55432336	rs2277339	0.11	G	4.80	7.14	5.36
12	111390798	rs11066320	0.35	A	12.00	10.92	9.26
12	113874813	rs12314403	0.09	G	5.94	7.90	6.32
17	70980029	rs6416847	0.18	G	3.69	7.51	5.76
19	48940718	rs346532	0.28	G	6.40	6.95	5.75

Table 4.10: **HaemgenRBC2012 NewSNPs** – Summary statistics of NewSNPs identified from the bmass analysis of HaemgenRBC2012. All phenotype ZScores are oriented towards the RBC minor allele (A1). Descriptions otherwise the same as previously detailed in GlobalLipids2010 NewSNPs (4.8).

Chr	BP	Marker	MAF	A1	unistat_log10pVal	mvstat_log10pVal	logBFWeightedAvg
2	19630595	rs1522484	0.30	T	5.03	7.20	5.88
7	2285786	rs2969070	0.35	G	8.59	6.90	6.30
12	111054406	rs17630235	0.42	A	11.53	9.76	9.11

Table 4.11: **ICBP2011 NewSNPs** – Summary statistics of NewSNPs identified from the bmass analysis of ICBP2011. All phenotype ZScores are oriented towards the SBP minor allele (A1). Descriptions otherwise the same as previously detailed in GlobalLipids2010 NewSNPs (4.8).

Chr	BP	Marker	MAF	A1	unistat_log10pVal	mvstat_log10pVal	logBFWeightedAvg
2	42291949	rs7560852	0.18	G	5.87	8.47	6.42
2	68997153	rs10196674	0.29	G	6.61	7.59	5.52
4	88750703	rs35481967	0.24	T	7.22	7.41	5.97
6	21391282	rs10946458	0.34	C	7.44	7.05	5.66
6	127144683	rs13194508	0.20	C	7.89	8.47	6.43
7	2394991	rs11551167	0.25	T	235.84	296.84	292.43
7	99130834	chr7:99130834	0.03	T	7.03	7.58	5.58
10	28478492	rs11006914	0.27	T	6.37	8.57	6.90
10	124088725	rs927427	0.45	C	6.41	6.81	5.33
11	16756873	rs11024028	0.15	G	7.84	8.93	7.61
11	86881464	rs570716	0.37	T	5.86	7.54	5.37
12	53664015	rs3741663	0.37	G	9.55	9.21	7.58
15	67414911	rs4776341	0.46	T	5.97	7.22	5.12
17	2048713	rs7209460	0.29	C	8.87	7.62	6.25
20	60990162	rs62196317	0.11	A	8.65	8.31	7.00

Table 4.12: **GEFOS2015 NewSNPs** – Summary statistics of NewSNPs identified from the bmass analysis of GEFOS2015. All phenotype ZScores are oriented towards the FA minor allele (A1). Descriptions otherwise the same as previously detailed in GlobalLipids2010 NewSNPs (4.8).

Chr	BP	Marker	MAF	A1	unistat_log10pVal	mvstat_log10pVal	logBFWeightedAvg
6	27145059	rs13194491	0.08	T	70.27	99.09	98.39
6	27694209	rs13217599	0.08	C	66.29	90.73	90.10
6	28893064	rs3118362	0.09	C	48.33	67.70	66.89
6	29618609	rs362538	0.08	C	36.54	45.89	44.85
6	30819784	rs3094124	0.09	C	17.28	22.25	20.58

Table 4.13: **GIS2014 NewSNPs** – Summary statistics of NewSNPs identified from the bmass analysis of GIS2014. All phenotype ZScores are oriented towards the Iron minor allele (A1). Descriptions otherwise the same as previously detailed in GlobalLipids2010 NewSNPs (4.8).

Chr	BP	Marker	MAF	A1	unistat_log10pVal	mvstat_log10pVal	logBFWeightedAvg
4	67806344	rs1969002	0.12	A	5.77	7.43	5.49

Table 4.14: **SSGAC2016 NewSNPs** – Summary statistics of NewSNPs identified from the bmass analysis of SSGAC2016. All phenotype ZScores are oriented towards the NEB_Pooled minor allele (A1). Descriptions otherwise the same as previously detailed in GlobalLipids2010 NewSNPs (4.8).

Chr	BP	Marker	MAF	A1	unistat_log10pVal	mvstat_log10pVal	logBFWeightedAvg
1	15717784	rs4233535	0.22	C	6.72	5.67	4.89
2	15780453	rs6431731	0.06	C	6.49	4.20	4.53
11	30705666	rs963837	0.47	C	7.28	5.09	5.30
12	15223609	rs2193172	0.11	C	5.77	6.24	4.26
17	19382628	rs2453583	0.43	T	6.24	4.17	4.29
17	34919080	rs12936996	0.24	G	6.55	4.66	4.63

Table 4.15: **CKDGen2010_1 NewSNPs** – Summary statistics of NewSNPs identified from the bmass analysis of CKDGen2010_1. All phenotype ZScores are oriented towards the Crea minor allele (A1). Descriptions otherwise the same as previously detailed in GlobalLipids2010 NewSNPs (4.8).

Chr	BP	Marker	MAF	A1	unistat_log10pVal	mvstat_log10pVal	logBFWeightedAvg
17	44837217	rs199529	0.24	C	6.78	8.39	8.34

Table 4.16: **EMERGE22015 NewSNPs** – Summary statistics of NewSNPs identified from the bmass analysis of EMERGE22015. All phenotype ZScores are oriented towards the ICV minor allele (A1). Descriptions otherwise the same as previously detailed in GlobalLipids2010 NewSNPs (4.8).

Chr	BP	Marker	MAF	A1	unistat_log10pVal	mvstat_log10pVal	logBFWeightedAvg
1	11797044	rs17367504	0.15	G	9.68	8.28	6.79
3	170583588	rs419076	0.47	T	8.52	7.13	6.10
3	170668995	rs1343040	0.43	A	8.65	7.01	6.15
4	81515724	rs13149993	0.32	A	10.48	10.99	9.57
4	81521902	rs1458038	0.28	T	12.10	11.55	10.61
5	32825609	rs1173756	0.46	T	7.39	6.15	5.35
6	26199158	rs1799945	0.15	G	7.32	6.17	5.24
6	26215442	rs198846	0.14	A	7.42	6.31	5.32
7	106005809	rs17477177	0.21	C	9.92	8.08	7.23
10	18747454	rs1813353	0.32	C	9.77	7.78	7.17
10	18767965	rs12258967	0.33	G	9.86	7.89	7.26
10	63137559	rs4590817	0.16	C	10.53	9.15	7.98
10	104585839	rs3824755	0.10	C	9.16	10.57	9.34
10	104836168	rs11191548	0.09	C	9.30	9.61	8.60
10	104929205	rs11191593	0.09	C	9.27	10.12	8.85
11	16858844	rs381815	0.26	T	8.61	9.11	8.08
11	129778440	rs11222084	0.35	T	10.07	9.08	8.44
12	88563054	rs17249754	0.16	A	12.01	11.42	9.78
12	110347328	rs3184504	0.48	T	13.63	11.63	10.86
12	110470476	rs653178	0.48	C	13.78	11.87	10.99
15	72864420	rs1378942	0.35	C	11.46	10.17	8.88
17	42368270	rs17608766	0.15	C	10.29	9.34	8.38

Table 4.17: **ICBP2011 PreviousSNPs** – Summary statistics of the previous univariate GWAS SNPs (PreviousSNPs) used for the bmass analysis of ICBP2011. All phenotype ZScores are oriented towards the SBP minor allele (A1). Descriptions otherwise the same as previously detailed in GlobalLipids2010 NewSNPs (4.8).

Chr	BP	Marker	MAF	A1	unistat_log10pVal	mvstat_log10pVal	logBFWeightedAvg
2	27594741	rs780094	0.39	T	11.60	11.37	10.31
2	169471394	rs560887	0.33	T	74.34	80.58	80.23
3	124548468	rs11708067	0.23	A	8.06	7.07	6.90
7	15030834	rs2191349	0.47	T	16.11	17.94	16.84
7	44202193	rs4607517	0.20	A	35.34	33.28	32.73
8	118254914	rs11558471	0.25	A	10.58	10.43	10.27
10	113032083	rs10885122	0.10	T	10.07	9.99	9.74
10	114746031	rs4506565	0.30	A	7.91	9.49	8.76
11	45829667	rs11605924	0.46	A	8.82	8.78	7.87
11	47292896	rs7944584	0.29	A	8.83	8.15	8.04
11	61328054	rs174550	0.37	T	7.83	7.57	7.21
11	92348358	rs10830963	0.30	C	67.90	71.01	70.34

Table 4.18: **MAGIC2010 PreviousSNPs** – Summary statistics of the previous univariate GWAS SNPs (PreviousSNPs) used for the bmass analysis of MAGIC2010. All phenotype ZScores are oriented towards the FstIns minor allele (A1). Descriptions otherwise the same as previously detailed in GlobalLipids2010 NewSNPs (4.8).

Chr	BP	Marker	MAF	A1	unistat_log10pVal	mvstat_log10pVal	logBFWeightedAvg
2	190086995	rs744653	0.16	C	12.92	15.21	13.95
3	134960391	rs8177240	0.36	G	308.43	320.63	321.59
3	197311602	rs9990333	0.42	T	10.52	10.78	9.42
6	26199158	rs1799945	0.13	G	59.53	58.21	56.98
6	26201120	rs1800562	0.04	A	177.82	241.64	241.35
8	18316746	rs4921915	0.25	G	10.76	9.32	8.50
11	61361390	rs174577	0.34	A	9.72	7.53	7.04
22	35792882	rs855791	0.39	A	79.46	79.94	78.48

Table 4.19: **GIS2014 PreviousSNPs** – Summary statistics of the previous univariate GWAS SNPs (PreviousSNPs) used for the bmass analysis of GIS2014. All phenotype ZScores are oriented towards the Iron minor allele (A1). Descriptions otherwise the same as previously detailed in GlobalLipids2010 NewSNPs (4.8).

Chr	BP	Marker	MAF	A1	unistat_log10pVal	mvstat_log10pVal	logBFWeightedAvg
1	153753725	rs10908474	0.28	A	7.72	11.10	9.00
1	153927052	rs10908557	0.23	G	8.76	10.88	8.79
2	100832218	rs1160544	0.39	A	8.22	8.17	6.40
3	49898000	rs2777888	0.45	A	14.42	14.23	12.29
5	45094503	rs6885307	0.18	C	9.65	8.83	7.47
6	152229850	rs2347867	0.30	G	8.83	10.85	8.77
7	114313218	rs10953766	0.42	A	9.78	8.87	7.57
20	31097877	rs293566	0.25	C	7.50	6.67	5.43
22	34503059	rs242997	0.44	G	8.37	7.64	6.26

Table 4.20: **SSGAC2016 PreviousSNPs** – Summary statistics of the previous univariate GWAS SNPs (PreviousSNPs) used for the bmass analysis of SSGAC2016. All phenotype ZScores are oriented towards the NEB_Pooled minor allele (A1). Descriptions otherwise the same as previously detailed in GlobalLipids2010 NewSNPs (4.8).

Chr	BP	Marker	MAF	A1	unistat_log10pVal	mvstat_log10pVal	logBFWeightedAvg
1	109801361	rs1933182	0.32	A	7.89	6.40	5.94
1	149218101	rs267734	0.26	C	8.28	7.13	6.24
2	5825331	rs16864170	0.07	C	7.35	5.55	4.10
2	27584444	rs1260326	0.40	T	9.89	9.54	7.95
2	73721836	rs13538	0.21	G	7.59	7.68	5.96
2	211248752	rs7422339	0.28	A	8.62	8.20	6.79
3	143289827	rs347685	0.25	C	8.15	6.40	6.19
4	77587871	rs17319721	0.46	A	18.96	24.64	22.70
5	39432889	rs11959928	0.43	A	10.74	9.88	8.66
5	176750242	rs6420094	0.42	G	11.42	9.66	9.29
6	43914587	rs881858	0.29	G	10.66	9.31	8.70
6	160588379	rs2279463	0.08	G	9.06	6.74	6.93
7	77254375	rs6465825	0.49	C	8.46	6.30	6.43
7	151038734	rs7805747	0.11	A	10.29	10.54	8.60
8	23807096	rs10109414	0.46	T	8.00	7.86	6.34
9	70624527	rs4744712	0.39	A	9.14	7.28	7.11
10	1146165	rs10794720	0.06	T	7.68	6.29	5.63
11	65263398	rs4014195	0.31	G	7.48	7.05	5.52
12	219559	rs10774021	0.33	C	8.17	6.84	6.23
12	110492139	rs653178	0.41	C	7.42	5.43	4.30
13	71245697	rs626277	0.44	C	9.54	7.34	7.45
15	43428517	rs2453533	0.29	A	21.34	19.89	18.85
15	51733885	rs491567	0.20	C	7.89	5.34	5.89
15	73946038	rs1394125	0.35	A	9.43	8.42	7.46
16	20275191	rs12917707	0.15	T	19.92	22.93	21.26
17	56811371	rs9895661	0.16	C	7.85	6.18	5.85
19	38048731	rs12460876	0.42	C	8.26	6.28	6.26
20	23560737	rs911119	0.26	C	137.64	136.79	138.40

Table 4.21: **CKDGen2010_1 PreviousSNPs** – Summary statistics of the previous univariate GWAS SNPs (PreviousSNPs) used for the bmass analysis of CKDGen2010_1. All phenotype ZScores are oriented towards the Crea minor allele (A1). Descriptions otherwise the same as previously detailed in GlobalLipids2010 NewSNPs (4.8).

Chr	BP	Marker	MAF	A1	unistat_log10pVal	mvstat_log10pVal	logBFWeightedAvg
12	65832468	rs61921502	0.14	G	9.53	8.41	8.48
12	117323367	rs77956314	0.08	C	10.03	9.08	9.55
17	43906828	rs17689882	0.23	A	8.54	9.64	9.97
18	50818827	rs62097986	0.39	A	10.32	9.60	10.64
20	30306724	rs6087771	0.33	C	8.62	7.84	7.48

Table 4.22: **EMERGE22015 PreviousSNPs** – Summary statistics of the previous univariate GWAS SNPs (PreviousSNPs) used for the bmass analysis of EMERGE22015. All phenotype ZScores are oriented towards the ICV minor allele (A1). Descriptions otherwise the same as previously detailed in GlobalLipids2010 NewSNPs (4.8).

	HDL_LDL_TG_TC	n	MeanPosterior	OriginalPrior
NewSNPs				
	1_1_1_2	18	0.557	0.381
	1_2_1_2	1	0.335	0.08
PreviousSNPs				
	1_1_1_2	56	0.595	0.381
	2_2_2_1	7	0.576	0.046
	1_2_2_2	6	0.324	0.036
	1_2_0_2	5	0.372	0.034
	2_1_1_0	5	0.654	0.046

Table 4.23: **GlobalLipids2010 Top Multivariate Models** – List of top multivariate models per NewSNPs and PreviousSNPs from GlobalLipids2010 analysis. See GlobalLipids2013 (4.5) for full description.

	Height_BMI_WHRadjBMI	n	MeanPosterior	OriginalPrior
NewSNPs				
	1_2_0	10	0.424	0.289
	1_1_0	3	0.811	0.173
	1_0_1	2	0.503	0.101
PreviousSNPs				
	1_2_0	81	0.43	0.289
	1_1_0	19	0.633	0.173
	1_0_1	16	0.482	0.101
	2_1_2	8	0.605	0.056
	1_1_1	4	0.487	0.037

Table 4.24: **GIANT2010 Top Multivariate Models** – List of top multivariate models per NewSNPs and PreviousSNPs from GIANT2010 analysis. See GlobalLipids2013 (4.5) for full description.

	Height_BMI_WHRadjBMI	n	MeanPosterior	OriginalPrior
NewSNPs				
	1_2_0	88	0.423	0.318
	1_1_1	21	0.656	0.161
	1_1_0	12	0.518	0.094
	1_2_1	1	0.336	0.037
PreviousSNPs				
	1_2_0	495	0.441	0.318
	1_1_1	114	0.68	0.161
	1_1_0	76	0.48	0.094
	1_2_1	14	0.391	0.037
	1_2_2	11	0.373	0.257

Table 4.25: **GIANT2014_5 Top Multivariate Models** – List of top multivariate models per NewSNPs and PreviousSNPs from GIANT2014_5 analysis. See GlobalLipids2013 (4.5) for full description.

	RBC_MCV_PCV MCH_Hb_MCHC	n	MeanPosterior	OriginalPrior
NewSNPs				
	2_1_0_1_2_1	5	0.499	0.116
	2_1_2_1_1_1	4	0.733	0.16
	1_1_2_2_2_2	2	0.475	0.074
	2_2_1_2_1_0	1	0.287	0.012
	1_1_2_1_2_1	1	0.515	0.081
PreviousSNPs				
	2_1_2_1_1_1	9	0.708	0.16
	2_1_0_1_2_1	9	0.567	0.116
	1_1_2_2_2_2	8	0.417	0.074
	2_1_0_2_0_0	5	0.308	0.038
	2_2_2_2_1_1	5	0.481	0.044

Table 4.26: **HaemgenRBC2012 Top Multivariate Models** – List of top multivariate models per NewSNPs and PreviousSNPs from HaemgenRBC2012 analysis. See GlobalLipids2013 (4.5) for full description.

	RBC_MCV_PCV MCH_Hb_MCHC	n	MeanPosterior	OriginalPrior
NewSNPs				
	2_1_2_2_1_1	28	0.486	0.203
	2_1_1_2_2_2	16	0.463	0.17
	2_1_0_2_2_1	9	0.415	0.117
	2_2_2_1_0_2	2	0.386	0.023
	2_1_1_2_2_1	2	0.353	0.155
PreviousSNPs				
	2_1_1_2_2_2	161	0.471	0.17
	2_1_2_2_1_1	131	0.471	0.203
	2_1_0_2_2_1	96	0.506	0.117
	2_1_1_2_2_1	49	0.569	0.155
	2_0_1_2_2_2	31	0.412	0.038

Table 4.27: **HaemgenRBC2016 Top Multivariate Models** – List of top multivariate models per NewSNPs and PreviousSNPs from HaemgenRBC2016 analysis. See GlobalLipids2013 (4.5) for full description.

	SBP_DBP_MAP_PP	n	MeanPosterior	OriginalPrior
NewSNPs				
	2_1_2_1	3	0.999	0.773
PreviousSNPs				
	2_1_2_1	17	0.967	0.773
	2_2_1_2	5	0.859	0.21

Table 4.28: **ICBP2011 Top Multivariate Models** – List of top multivariate models per NewSNPs and PreviousSNPs from ICBP2011 analysis. See GlobalLipids2013 (4.5) for full description.

	FstIns_FstGlu HOMA_B_HOMA_IR	n	MeanPosterior	OriginalPrior
PreviousSNPs				
	2_1_2_0	8	0.795	0.588
	0_1_2_2	2	0.834	0.236
	2_1_0_2	1	0.541	0.045
	2_1_1_2	1	0.519	0.131

Table 4.29: **MAGIC2010 Top Multivariate Models** – List of top multivariate models per PreviousSNPs from MAGIC2010 analysis. See GlobalLipids2013 (4.5) for full description.

	FA_FN_LS	n	MeanPosterior	OriginalPrior
NewSNPs				
	2_1_1	8	0.782	0.425
	1_1_1	7	0.872	0.481
PreviousSNPs				
	1_1_1	17	0.773	0.481
	2_1_1	14	0.783	0.425
	1_2_1	3	0.885	0.094

Table 4.30: **GEFOS2015 Top Multivariate Models** – List of top multivariate models per NewSNPs and PreviousSNPs from GEFOS2015 analysis. See GlobalLipids2013 (4.5) for full description.

	Iron_Sat_TrnsFrn_Log10Frtn	n	MeanPosterior	OriginalPrior
NewSNPs				
	2_1_1_2	5	0.995	0.257
PreviousSNPs				
	2_1_1_2	2	0.978	0.257
	0_2_1_2	1	0.759	0.14
	1_1_2_2	1	0.977	0.125
	0_2_1_1	1	0.999	0.164
	1_1_1_1	1	1	0.129

Table 4.31: **GIS2014 Top Multivariate Models** – List of top multivariate models per NewSNPs and PreviousSNPs from GIS2014 analysis. See GlobalLipids2013 (4.5) for full description.

	NEB_Pooled_AFB_Pooled	n	MeanPosterior	OriginalPrior
NewSNPs				
	1_1	1	0.983	.558
PreviousSNPs				
	1_1	5	0.827	0.558
	2_1	4	0.778	0.442

Table 4.32: **SSGAC2016 Top Multivariate Models** – List of top multivariate models per NewSNPs and PreviousSNPs from SSGAC2016 analysis. See Global-Lipids2013 (4.5) for full description.

	Crea_Cys_CKD_UACR_MA	n	MeanPosterior	OriginalPrior
NewSNPs				
	1_2_2_2_0	6	0.6	0.473
PreviousSNPs				
	1_2_2_2_0	20	0.625	0.473
	1_1_2_2_1	3	0.61	0.108
	1_1_1_0_0	2	0.536	0.057
	1_1_2_0_2	1	0.446	0.056
	0_1_2_2_0	1	0.999	0.038

Table 4.33: **CKDGen2010_1 Top Multivariate Models** – List of top multivariate models per NewSNPs and PreviousSNPs from CKDGen2010_1 analysis. See GlobalLipids2013 (4.5) for full description.

	ICV_Accumbens_Amygdala_Caudate Hippocampus_Pallidum_Putamen_Thalamus	n	MeanPosterior	OriginalPrior
NewSNPs				
	1_0_0_0_0_0_1_1	1	0.692	0.148
PreviousSNPs				
	0_0_0_0_1_0_0_0	1	0.841	0.169
	0_2_2_2_1_1_0_2	1	0.903	0.182
	1_0_0_0_0_0_1_1	1	0.701	0.148
	0_0_2_0_0_2_1_2	1	0.893	0.179
	0_2_0_1_0_2_1_2	1	0.997	0.208

Table 4.34: **EMERGE22015 Top Multivariate Models** – List of top multivariate models per NewSNPs and PreviousSNPs from EMERGE22015 analysis. See GlobalLipids2013 (4.5) for full description.

CHAPTER 5

CONCLUSIONS

A major goal of human genetics research is to connect naturally occurring genotyping and phenotypic variation. With limitations on the types of experimental procedures human geneticists can conduct, finding associations between DNA markers and phenotypic outcomes of interest is a long-standing theme in the research we conduct. Recent advances in both genotyping and sequencing technologies have made it significantly more affordable to collect appreciable amounts of genetic information from large cohorts of individuals. But with these new technologies and increasing amounts of data come new challenges. Oftentimes researchers are quick to try and answer these challenges by simply collecting more data and/or improving technologies. While these are reasonable approaches, they should be complementary to efforts focused on better understanding the data we already have. Developing improved models to explain biological phenomenon and methods to parse the data we have is just as important to face the challenges these new technological advancements contain. Here we present three projects that aim to overcome pre-existing challenges in human genetics by using data beyond just an initial first-pass analysis. In particular, all three projects use summary information to either repurpose existing data or better inform follow-up analyses.

In Chapter 2 we investigated whether recent demographic changes in human history have differentially impacted the deleterious mutational load in two human popula-

tions. With the advent of more affordable sequencing technologies, recent large-scale investigations of the human genome have revealed a substantial excess of rare variants in many human populations[34, 148, 68, 112, 158, 215]. This is thought to be due to the recent explosive population growth humans have experienced. These studies also observed greater proportions of rare variants in European and Asian populations compared to African populations, possibly due to the ancient out-of-Africa bottleneck the former populations experienced[233, 229, 113, 84, 215]. Therefore we explored whether these varying demographic histories, and their apparent impacts on rare variants, would produce observable difference in the deleterious mutational load between human populations. First, we used simulations and theoretical predictions to show that for most biological scenarios (e.g. different levels of selection and dominance) we do not expect to see a difference in the mutational load between populations. We also showed theoretically that for most ranges of selection we do not expect rare variants to substantially affect the genetic variation of complex traits. Second, we used allele frequency summary information from two datasets (the Exome Sequencing Project[68] and 1000Genomes[217]) to directly calculate the observed deleterious mutational load in both European-Americans and African-Americans. Using functional prediction algorithms[6, 121, 32, 191] to designate whether variants were deleterious or not, we found that on both the population-level as well as on the individual-level, the deleterious mutational load does not significantly differ between populations. Therefore overall we showed that despite differences in the recent demographic histories of human populations, a) we should not expect to see major differences in the mutational burden between these populations, and b) in-

deed that is what we observe using real exome data from European-Americans and African-Americans from two different datasets.

One interesting result from this project pertains to the affects of newer technologies on population genetics analyses. As mentioned, the questions being addressed here are not novel; previous studies have attempted to answer these questions, but had access either to fewer genomic regions[34, 158] or smaller sample sizes[113, 139]. But with the availability of whole-exome data, we were able to address this question using both all genes present in the genome as well as larger sample sizes. It is possible that some of the discrepancies presented here between our results and earlier work are due to changes in technologies (such as having an agnostic vs. targeted collection of loci). In fact recent work in the field of human *de novo* mutation rates has addressed this very concern. Historically the yearly human germline mutation rate has been estimated as 1.0×10^{-9} per bp, but since the rise and analysis of whole-genome sequencing, this yearly rate has now been revised to $.5\times10^{-9}$ per bp (reviewed and discussed by Moorjani et al. 2016[154]). Therefore using newer technologies can sometimes reveal important changes to long-standing assumptions, and our study here provides the most technologically up-to-date (at the time) analysis of the human deleterious mutation load. Additionally, previous work in this field has often addressed the impact of rare variation on complex traits through the use of the functional prediction algorithms; studies would imply that an abundance of rare, deleterious mutations in the exomes of genes should affect complex traits. However in our study we provide direct modeling of complex trait architecture in an attempt to answer this question more thoroughly.

In Chapter 3 we presented the first stage results of a two-phase HIV candidate-gene exome sequencing study. Infection with HIV is an ongoing world-wide epidemic, affecting more than 30 million individuals currently[160, 142, 1]. Past human genetics research has identified many promising candidate genes[94, 9, 192], but after a number of initial GWAS few regions outside HLA achieved significant levels of association using various HIV-related phenotypes (e.g. presence o absence of HIV, levels of $CD4^+$ cell counts) [57, 24, 58, 169, 151]. In response to these ambiguous results, a group of researchers – the HIV Immune Network Team (HINT) – aimed to develop a more targeted approach for identifying human genes functionally related to HIV, as well as to explore the impact of rare variation on HIV-susceptibility. Using a number of in-house and publicly available experiments, HINT researchers created a list of ∼1,700 genes that appeared to have strong *a priori* evidence for being related to HIV. They then created custom arrays to target the exomes of these genes and subsequently sequenced them in a subset of individuals from a commonly-used HIV cohort (Multicenter AIDS Cohort Study, MACS[110]). Here, we show the results from our bioinformatics pipeline to process this raw sequencing data as well as the analytical results from conducting gene-level association analyses. Specifically, we discovered 149,063 high-quality variants across ∼1,300 genes that include 8,842 non-synonymous SNPs. We then conducted gene-level analyses using SKAT-O, a second-generation rare-variant test, to identify whether any genes were significantly associated with HIV-Acquisition (HIV seronegative vs. HIV seropositive individuals) or AIDS progression (very rapid + rapid vs. very slow + slow AIDS progressors). To test genes, we used all the discovered non-synonymous variants for a given locus. We

found an enrichment of marginally significant associations in HIV-acquisition when comparing our results to test statistics from 1,000 permutations of the data. We then used these results to help inform and design a custom follow-up genotyping array, which will soon be run on the full MACS cohort. Over 29,000 variants have been included on this follow-up genotyping array, which included all the non-synonymous variants we discovered as well as variants in top genes from external resources such as ExAC[131] and GTEX[33]. Overall, we showed the initial association results from the first stage of this candidate gene-exome study, and the steps being taken to move this project into its second phase.

One important question still being addressed by this study is the impact of rare variation on HIV-related phenotypes. As previously mentioned, the initial GWAS conducted in HIV were mainly inconclusive, thus suggesting common variation may not play an important role in HIV-susceptibility. Since HIV is a relatively recent infectious disease, it may make sense that rare variation, alternatively, is of particular importance; HIV may have evolved by hijacking newly arisen mutations fluctuating in human populations. Recent work has suggested much of the human genome has already faced selective pressures from past viruses[50], perhaps indicating that newer viruses may find more novel host-vulnerabilities among rare, recent mutations instead. And indeed our current results suggest rare variation may play a role in HIV-susceptibility; our SKAT-O analyses produce an excess of marginally significant associations between genes and HIV-acquisition. However, no single gene reaches genome-wide significance; this could be due to our modest sample size (<1000) or because rare variation does in fact have a limited impact on HIV-acquisition. Ex-

panding our analysis to the full MACS cohort should differentiate between these two scenarios and provide a clearer answer for the human genetics community.

Additionally, another open question our study deals with is how best to design a rare-variant association study. As newer technologies such as whole-exome and whole-genome sequencing become more affordable, it is fair to ask whether targeted approaches as done here are still applicable. Presumably targeted approaches should still be more cost effective and in turn allow a larger number of samples processed per project budget. Additionally for traits where rare variation plays a greater role, it may be particularly important to sequence a larger number of individuals – with each new individual, another previously undiscovered rare variant might be captured. However, this strategy is only as powerful as your prior evidence for choosing loci. Perhaps doing an initial round of whole-exome or whole-genome sequencing will ultimately become a universally better choice, and targeted analysis will be more suitable as a follow-up sequencing or genotyping role. Therefore with the second stage of this study we aim to provide the community with answers regarding not only the biology of HIV-infection but also the efficacy of our design strategy.

In Chapter 4 we presented the results from applying a Bayesian multivariate GWAS method (bmass) on 13 publicly available datasets, providing both many new results as well as the phenotypic patterns driving these results. It is well appreciated by this point that using multivariate methods in GWAS can increase the power to detect signals of association[105, 247, 196, 244, 72]. However, despite this, multivariate GWAS methods are not frequently used, even when published studies contain mul-

tiple phenotypes. Here, we extended a previously published Bayesian multivariate framework[207] by making it into a readily usable R package. Bayesian multivariate analysis of summary statistics, or bmass, runs on univariate GWAS summary statistics, can quickly analyze datasets containing up to 8 phenotypes, and returns both new significant associations as well as the top multivariate patterns found in the data. We analyzed a diverse set of publicly available studies, containing a range of sample sizes genotyped and phenotypes measured. For a number of the datasets included we found many new associations, including over 50 novel loci for multiple studies. For all datasets analyzed we also provided the top multivariate models for each new association as well as the broad phenotypic patterns seen across entire datasets. We also showed how you can use differing phenotypic patterns to better refine association signals between nearby SNPs. Lastly, we highlighted particularly interesting signals of association, such as rs11708067 and *ADCY5* – a locus that has been connected to both hereditary motor function dysregulation disorders[31, 30] as well as metabolic traits[48, 188, 179, 91] – being significantly associated with both Height and BMI.

Overall through this project we have provided the human genetics community with multiple examples of multivariate GWAS analysis as well as software to efficiently conduct multivariate GWAS. Specifically we provide one of the most phenotypically complex examples of multivariate GWAS to date. Previously Pickrell et. al 2016 represented one of the most expansive studies of multivariate analysis in GWAS; however while their work provided many examples of pleiotropic SNPs across a wide range of phenotypes, they were limited to only analyzing bivariate models of as-

sociation. Because we are able to analyze up to 8 phenotypes through bmass, we provide a more extensive analysis of the multivariate space for a given SNP association. Additionally, we provide the community with software that a) explicitly tests every possible model and b) presents results as Bayes Factors. Unlike other multivariate GWAS programs that explore either more restricted or ambiguous sets of models[147, 115, 62, 162, 224], bmass allows for better interpretation of results by providing the full context of the multivariate space.

We also attempt to provide the community with a multivariate model that distinguishes between direct associations and indirect associations. Interestingly, we find that properly assigning indirect associations to be a challenge. At times there appears to be an excess of phenotypes falling into the indirect category across SNP associations; we believe this partially may be due to the indirect category being a 'default' category of sorts. If a phenotype does not fit into the **U** or **D** categories for a given SNP (which leads to an extra degree of freedom being included in the test for association), it may be 'easier' for the phenotype to be assigned to the **I** category (which effectively removes the phenotype from being included in the test for association, thus incurring no additional penalty). We think this may be due to noisy beta estimates (larger standard errors) being treated equally to more precise beta estimates (smaller standard errors) – this extra noise may be obscuring a phenotype's proper assignment to either **U** or **D**. One way to deal with this issue may be to shrink these poor estimates and lessen their impact on model assignment, and recent work in the Stephens Lab has focused on this issue.

Specifically, the Stephens Lab has been developing a framework known as Adaptive SHrinkage ('ash') [208]. ash attempts to accomplish this aforementioned shrinkage through one key advancement: using two summary statistics instead of one. In lieu of the typical single summary statistic that most GWAS-correction methods use (i.e. Z-scores or p-values), ash utilizes from a given association analysis both the beta estimates and their respective standard errors. By using two summary statistics versus one, there is a greater ability to estimate variance in measurement precision, which in turns allows for more accurate shrinkage of imprecise estimates. Additionally, by using two summary statistics, it is possible to calculate a 'local false sign rate' per beta estimate – confidence intervals for the sign of an effect size and not just whether it is non-zero. This 'local false sign rate' has multiple benefits, including more fine-grained evaluation of individual results and robustness to modeling assumptions. The Stephens Lab has been actively applying ash to multiple research directions, including the analysis of multiple traits in association studies[220]. The current application of ash to multivariate GWAS however lacks the phenotypic modeling presented here, i.e. the $\{\mathbf{U}, \mathbf{D}, \mathbf{I}\}$ designations. Therefore future work for this project includes combining ash with the bmass-modeling framework; doing so may then allow us to better assign phenotypes to their true multivariate categories, such as direct versus indirection associations.

5.1 Concluding Remarks

In this dissertation, we aimed to show working with data beyond first-pass analyses can produce more interesting and beneficial results. The field of human genetics is at an exciting time with the growing spectrum of datasets, technologies, and intermediate phenotypes that are now available. The work presented here is not meant to downplay the importance of these developments or dismiss the potential resources such as BioBanks and technologies such as single-cell sequencing contain. However, there are limits to what adding more data or using newer technology can accomplish. If for example epistatic interactions are essential to understanding all genetic mechanisms, it will not matter how many more genomes we have access to if we do not properly develop models and methods to accurately identify epistatic interactions. Ultimately, adding new resources should be one of multiple efforts being equally pursued to solve ongoing challenges in human genetics. While some of the problems laid out by "The Missing Heritability" papers[144, 200, 145, 249] have evolved and changed since their publications, one of the core messages remains the same: we should take the time to think more deeply about the data and results we have currently generated, and how best to utilize them.

References

[1] 2014(August 7th, 2014.).

[2] Kaposi's sarcoma and pneumocystis pneumonia among homosexual men–new york city and california. *MMWR Morb Mortal Wkly Rep*, 30(25):305–8, 1981.

[3] Pneumocystis pneumonia–los angeles. *MMWR Morb Mortal Wkly Rep*, 30(21):250–2, 1981.

[4] G. R. Abecasis, A. Auton, L. D. Brooks, M. A. DePristo, R. M. Durbin, R. E. Handsaker, H. M. Kang, G. T. Marth, and G. A. McVean. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, 2012.

[5] I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, and S. R. Sunyaev. A method and server for predicting damaging missense mutations. *Nat Methods*, 7(4):248–9, 2010.

[6] Ivan A Adzhubei, Steffen Schmidt, Leonid Peshkin, Vasily E Ramensky, Anna Gerasimova, Peer Bork, Alexey S Kondrashov, and Shamil R Sunyaev. A method and server for predicting damaging missense mutations. *Nature Methods*, 7(4):248–249, 2010.

[7] O. S. Ahmad, J. A. Morris, M. Mujammami, V. Forgetta, A. Leong, R. Li, M. Turgeon, C. M. Greenwood, G. Thanassoulis, J. B. Meigs, R. Sladek, and J. B. Richards. A mendelian randomization study of the effect of type-2 diabetes on coronary heart disease. *Nat Commun*, 6:7060, 2015.

[8] F. W. Albert and L. Kruglyak. The role of regulatory variation in complex traits and disease. *Nat Rev Genet*, 16(4):197–212, 2015.

[9] P. An and C. A. Winkler. Host genes associated with hiv/aids: advances in gene discovery. *Trends Genet*, 26(3):119–31, 2010.

[10] E. A. Andersson, K. Pilgaard, C. Pisinger, M. N. Harder, N. Grarup, K. Faerch, P. Poulsen, D. R. Witte, T. Jorgensen, A. Vaag, T. Hansen, and O. Pedersen. Type 2 diabetes risk alleles near adcy5, cdkal1 and hhex-ide are associated with reduced birthweight. *Diabetologia*, 53(9):1908–16, 2010.

[11] W. J. Astle, H. Elding, T. Jiang, D. Allen, D. Ruklisa, A. L. Mann, D. Mead, H. Bouman, F. Riveros-Mckay, M. A. Kostadima, J. J. Lambourne, S. Sivapalaratnam, K. Downes, K. Kundu, L. Bomba, K. Berentsen, J. R. Bradley, L. C. Daugherty, O. Delaneau, K. Freson, S. F. Garner, L. Grassi, J. Guerrero, M. Haimel, E. M. Janssen-Megens, A. Kaan, M. Kamat, B. Kim, A. Mandoli, J. Marchini, J. H. Martens, S. Meacham, K. Megy, J. O'Connell, R. Petersen, N. Sharifi, S. M. Sheard, J. R. Staley, S. Tuna, M. van der Ent, K. Walter, S. Y. Wang, E. Wheeler, S. P. Wilder, V. Iotchkova, C. Moore, J. Sambrook, H. G. Stunnenberg, E. Di Angelantonio, S. Kaptoge, T. W. Kuijpers, E. Carrillo-de Santa-Pau, D. Juan, D. Rico, A. Valencia, L. Chen, B. Ge, L. Vasquez, T. Kwan, D. Garrido-Martin, S. Watt, Y. Yang, R. Guigo, S. Beck, D. S. Paul, T. Pastinen, D. Bujold, G. Bourque, M. Frontini, J. Danesh, D. J. Roberts, W. H. Ouwehand, A. S. Butterworth, and N. Soranzo. The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell*, 167(5):1415–1429 e19, 2016.

[12] N. Barban, R. Jansen, R. de Vlaming, A. Vaez, J. J. Mandemakers, F. C. Tropf, X. Shen, J. F. Wilson, D. I. Chasman, I. M. Nolte, V. Tragante, S. W. van der Laan, J. R. Perry, A. Kong, Bios Consortium, T. S. Ahluwalia, E. Albrecht, L. Yerges-Armstrong, G. Atzmon, K. Auro, K. Ayers, A. Bakshi, D. Ben-Avraham, K. Berger, A. Bergman, L. Bertram, L. F. Bielak, G. Bjornsdottir, M. J. Bonder, L. Broer, M. Bui, C. Barbieri, A. Cavadino, J. E. Chavarro, C. Turman, M. P. Concas, H. J. Cordell, G. Davies, P. Eibich, N. Eriksson, T. Esko, J. Eriksson, F. Falahi, J. F. Felix, M. A. Fontana, L. Franke, I. Gandin, A. J. Gaskins, C. Gieger, E. P. Gunderson, X. Guo, C. Hayward, C. He, E. Hofer, H. Huang, P. K. Joshi, S. Kanoni, R. Karlsson, S. Kiechl, A. Kifley, A. Kluttig, P. Kraft, V. Lagou, C. Lecoeur, J. Lahti, R. Li-Gao, P. A. Lind, T. Liu, E. Makalic, C. Mamasoula, L. Matteson, H. Mbarek, P. F. McArdle, G. McMahon, S. F. Meddens, E. Mihailov, M. Miller, S. A. Missmer, C. Monnereau, P. J. van der Most, R. Myhre, M. A. Nalls, T. Nutile, I. P. Kalafati, E. Porcu, I. Prokopenko, K. B. Rajan, J. Rich-Edwards, C. A. Rietveld, A. Robino, L. M. Rose, R. Rueedi, K. A. Ryan, Y. Saba, D. Schmidt, J. A. Smith, L. Stolk, E. Streeten, A. Tonjes, G. Thorleifsson, et al. Genome-wide analysis identifies 12 loci influencing human reproductive behavior. *Nat Genet*, 48(12):1462–1472, 2016.

[13] J. C. Barrett and L. R. Cardon. Evaluating coverage of genome-wide association studies. *Nat Genet*, 38(6):659–62, 2006.

[14] J. C. Barrett, S. Hansoul, D. L. Nicolae, J. H. Cho, R. H. Duerr, J. D. Rioux, S. R. Brant, M. S. Silverberg, K. D. Taylor, M. M. Barmada, A. Bitton, T. Dassopoulos, L. W. Datta, T. Green, A. M. Griffiths, E. O. Kistner, M. T. Murtha, M. D. Regueiro, J. I. Rotter, L. P. Schumm, A. H. Steinhart, S. R. Targan, R. J. Xavier, Niddk Ibd Genetics Consortium, C. Libioulle, C. Sandor, M. Lathrop, J. Belaiche, O. Dewit, I. Gut, S. Heath, D. Laukens, M. Mni, P. Rutgeerts, A. Van Gossum, D. Zelenika, D. Franchimont, J. P. Hugot, M. de Vos, S. Vermeire, E. Louis, I. B. D. Consortium Belgian-French, Consortium Wellcome Trust Case Control, L. R. Cardon, C. A. Anderson, H. Drummond, E. Nimmo, T. Ahmad, N. J. Prescott, C. M. Onnie, S. A. Fisher, J. Marchini, J. Ghori, S. Bumpstead, R. Gwilliam, M. Tremelling, P. Deloukas, J. Mansfield, D. Jewell, J. Satsangi, C. G. Mathew, M. Parkes, M. Georges, and M. J. Daly. Genome-wide association defines more than 30 distinct susceptibility loci for crohn's disease. *Nat Genet*, 40(8):955–62, 2008.

[15] M. Beaudoin, P. Goyette, G. Boucher, K. S. Lo, M. A. Rivas, C. Stevens, A. Alikashani, M. Ladouceur, D. Ellinghaus, L. Torkvist, G. Goel, C. Lagace, V. Annese, A. Bitton, J. Begun, S. R. Brant, F. Bresso, J. H. Cho, R. H. Duerr, J. Halfvarson, D. P. McGovern, G. Radford-Smith, S. Schreiber, P. L. Schumm, Y. Sharma, M. S. Silverberg, R. K. Weersma, I. B. D. Genetics Consortium Quebec, Niddk Ibd Genetics Consortium, I. B. D. Genetics Consortium International, M. D'Amato, S. Vermeire, A. Franke, G. Lettre, R. J. Xavier, M. J. Daly, and J. D. Rioux. Deep resequencing of gwas loci identifies rare variants in card9, il23r and rnf186 that are associated with ulcerative colitis. *PLoS Genet*, 9(9):e1003723, 2013.

[16] B. Benyamin, T. Esko, J. S. Ried, A. Radhakrishnan, S. H. Vermeulen, M. Traglia, M. Gogele, D. Anderson, L. Broer, C. Podmore, J. Luan, Z. Kutalik, S. Sanna, P. van der Meer, T. Tanaka, F. Wang, H. J. Westra, L. Franke, E. Mihailov, L. Milani, J. Halldin, J. Winkelmann, T. Meitinger, J. Thiery, A. Peters, M. Waldenberger, A. Rendon, J. Jolley, J. Sambrook, L. A. Kiemeney, F. C. Sweep, C. F. Sala, C. Schwienbacher, I. Pichler, J. Hui, A. Demirkan, A. Isaacs, N. Amin, M. Steri, G. Waeber, N. Verweij, J. E. Powell, D. R. Nyholt, A. C. Heath, P. A. Madden, P. M. Visscher, M. J. Wright, G. W. Montgomery, N. G. Martin, D. Hernandez, S. Bandinelli, P. van der Harst, M. Uda, P. Vollenweider, R. A. Scott, C. Langenberg, N. J. Wareham, Consortium InterAct, C. van Duijn, J. Beilby, P. P. Pramstaller, A. A. Hicks, W. H. Ouwehand, K. Oexle, C. Gieger, A. Metspalu, C. Camaschella, D. Toniolo, D. W. Swinkels, and J. B. Whitfield. Novel loci affecting iron

homeostasis and their effects in individuals at risk for hemochromatosis. *Nat Commun*, 5:4926, 2014.

[17] E. Birney, G. D. Smith, and J. M. Greally. Epigenome-wide association studies and the interpretation of disease -omics. *PLoS Genet*, 12(6):e1006105, 2016.

[18] J. A. Blake, J. T. Eppig, J. A. Kadin, J. E. Richardson, C. L. Smith, C. J. Bult, and Group the Mouse Genome Database. Mouse genome database (mgd)-2017: community knowledge resource for the laboratory mouse. *Nucleic Acids Res*, 45(D1):D723–D729, 2017.

[19] C. A. Boger, M. H. Chen, A. Tin, M. Olden, A. Kottgen, I. H. de Boer, C. Fuchsberger, C. M. O'Seaghdha, C. Pattaro, A. Teumer, C. T. Liu, N. L. Glazer, M. Li, J. R. O'Connell, T. Tanaka, C. A. Peralta, Z. Kutalik, J. Luan, J. H. Zhao, S. J. Hwang, E. Akylbekova, H. Kramer, P. van der Harst, A. V. Smith, K. Lohman, M. de Andrade, C. Hayward, B. Kollerits, A. Tonjes, T. Aspelund, E. Ingelsson, G. Eiriksdottir, L. J. Launer, T. B. Harris, A. R. Shuldiner, B. D. Mitchell, D. E. Arking, N. Franceschini, E. Boerwinkle, J. Egan, D. Hernandez, M. Reilly, R. R. Townsend, T. Lumley, D. S. Siscovick, B. M. Psaty, B. Kestenbaum, T. Haritunians, S. Bergmann, P. Vollenweider, G. Waeber, V. Mooser, D. Waterworth, A. D. Johnson, J. C. Florez, J. B. Meigs, X. Lu, S. T. Turner, E. J. Atkinson, T. S. Leak, K. Aasarod, F. Skorpen, A. C. Syvanen, T. Illig, J. Baumert, W. Koenig, B. K. Kramer, O. Devuyst, J. C. Mychaleckyj, C. Minelli, S. J. Bakker, L. Kedenko, B. Paulweber, S. Coassin, K. Endlich, H. K. Kroemer, R. Biffar, S. Stracke, H. Volzke, M. Stumvoll, R. Magi, H. Campbell, V. Vitart, N. D. Hastie, V. Gudnason, S. L. Kardia, Y. Liu, O. Polasek, G. Curhan, F. Kronenberg, I. Prokopenko, I. Rudan, J. Arnlov, S. Hallan, G. Navis, C. KDGen Consortium, A. Parsa, L. Ferrucci, J. Coresh, M. G. Shlipak, et al. Cubn is a gene locus for albuminuria. *J Am Soc Nephrol*, 22(3):555–70, 2011.

[20] D. Botstein, R. L. White, M. Skolnick, and R. W. Davis. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet*, 32(3):314–31, 1980.

[21] E. A. Boyle, Y. I. Li, and J. K. Pritchard. An expanded view of complex traits: From polygenic to omnigenic. *Cell*, 169(7):1177–1186, 2017.

[22] A. L. Brass, D. M. Dykxhoorn, Y. Benita, N. Yan, A. Engelman, R. J. Xavier, J. Lieberman, and S. J. Elledge. Identification of host proteins required for hiv infection through a functional genomic screen. *Science*, 319(5865):921–6, 2008.

[23] S. Burgess, N. J. Timpson, S. Ebrahim, and G. Davey Smith. Mendelian randomization: where are we now and where are we going? *Int J Epidemiol*, 44(2):379–88, 2015.

[24] F. D. Bushman, N. Malani, J. Fernandes, I. D'Orso, G. Cagney, T. L. Diamond, H. Zhou, D. J. Hazuda, A. S. Espeseth, R. Konig, S. Bandyopadhyay, T. Ideker, S. P. Goff, N. J. Krogan, A. D. Frankel, J. A. Young, and S. K. Chanda. Host cell factors in hiv replication: meta-analysis of genome-wide studies. *PLoS Pathog*, 5(5):e1000437, 2009.

[25] C. J. Cardinale, J. R. Kelsen, R. N. Baldassano, and H. Hakonarson. Impact of exome sequencing in inflammatory bowel disease. *World J Gastroenterol*, 19(40):6721–9, 2013.

[26] L. R. Cardon and J. I. Bell. Association study designs for complex diseases. *Nat Rev Genet*, 2(2):91–9, 2001.

[27] C. S. Carlson, M. A. Eberle, M. J. Rieder, J. D. Smith, L. Kruglyak, and D. A. Nickerson. Additional snps and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans. *Nat Genet*, 33(4):518–21, 2003.

[28] F. Casals and J. Bertranpetit. Human genetic variation, shared and private. *Science*, 337(6090):39–40, 2012.

[29] B Charlesworth and D Charlesworth. Elements of evolutionary genetics., 2010.

[30] Y. Z. Chen, J. R. Friedman, D. H. Chen, G. C. Chan, C. S. Bloss, F. M. Hisama, S. E. Topol, A. R. Carson, P. H. Pham, E. S. Bonkowski, E. R. Scott, J. K. Lee, G. Zhang, G. Oliveira, J. Xu, A. A. Scott-Van Zeeland, Q. Chen, S. Levy, E. J. Topol, D. Storm, P. D. Swanson, T. D. Bird, N. J. Schork, W. H. Raskind, and A. Torkamani. Gain-of-function adcy5 mutations in familial dyskinesia with facial myokymia. *Ann Neurol*, 75(4):542–9, 2014.

[31] Y. Z. Chen, M. M. Matsushita, P. Robertson, M. Rieder, S. Girirajan, F. Antonacci, H. Lipe, E. E. Eichler, D. A. Nickerson, T. D. Bird, and W. H. Raskind. Autosomal dominant familial dyskinesia and facial myokymia: single exome sequencing identifies a mutation in adenylyl cyclase 5. *Arch Neurol*, 69(5):630–5, 2012.

[32] Sung Chun and Justin C Fay. Identification of deleterious mutations within three human genomes. *Genome research*, 19(9):1553–1561, 2009.

[33] G. TEx Consortium. The genotype-tissue expression (gtex) project. *Nat Genet*, 45(6):580–5, 2013.

[34] A. Coventry, L.M. Bull-Otterson, X. Liu, A.G. Clark, T.J. Maxwell, J. Crosby, J.E. Hixson, T.J. Rea, D.M. Muzny, L.R. Lewis, et al. Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nature Communications*, 1:131, 2010.

[35] M. J. Daly, J. D. Rioux, S. F. Schaffner, T. J. Hudson, and E. S. Lander. High-resolution haplotype structure in the human genome. *Nat Genet*, 29(2):229–32, 2001.

[36] P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, R. Durbin, and Group Genomes Project Analysis. The variant call format and vcftools. *Bioinformatics*, 27(15):2156–8, 2011.

[37] A. Darvasi and M. Soller. Advanced intercross lines, an experimental population for fine genetic mapping. *Genetics*, 141(3):1199–207, 1995.

[38] J. R. David, P. Gibert, H. Legout, G. Petavy, P. Capy, and B. Moreteau. Isofemale lines in drosophila: an empirical approach to quantitative trait analysis in natural populations. *Heredity (Edinb)*, 94(1):3–12, 2005.

[39] K. E. Davies. The application of dna recombinant technology to the analysis of the human genome and genetic disease. *Hum Genet*, 58(4):351–7, 1981.

[40] T. A. Dayeh, A. H. Olsson, P. Volkov, P. Almgren, T. Ronn, and C. Ling. Identification of cpg-snps associated with type 2 diabetes and differential dna methylation in human pancreatic islets. *Diabetologia*, 56(5):1036–46, 2013.

[41] S. De Rubeis, X. He, A. P. Goldberg, C. S. Poultney, K. Samocha, A. E. Cicek, Y. Kou, L. Liu, M. Fromer, S. Walker, T. Singh, L. Klei, J. Kosmicki, F. Shih-Chen, B. Aleksic, M. Biscaldi, P. F. Bolton, J. M. Brownfeld, J. Cai, N. G. Campbell, A. Carracedo, M. H. Chahrour, A. G. Chiocchetti, H. Coon, E. L. Crawford, S. R. Curran, G. Dawson, E. Duketis, B. A. Fernandez, L. Gallagher, E. Geller, S. J. Guter, R. S. Hill, J. Ionita-Laza, P. Jimenz Gonzalez, H. Kilpinen, S. M. Klauck, A. Kolevzon, I. Lee, I. Lei, J. Lei, T. Lehtimaki, C. F. Lin, A. Ma'ayan, C. R. Marshall, A. L. McInnes, B. Neale, M. J. Owen, N. Ozaki, M. Parellada, J. R. Parr, S. Purcell, K. Puura, D. Rajagopalan, K. Rehnstrom, A. Reichenberg, A. Sabo, M. Sachse, S. J. Sanders, C. Schafer,

M. Schulte-Ruther, D. Skuse, C. Stevens, P. Szatmari, K. Tammimies, O. Valladares, A. Voran, W. Li-San, L. A. Weiss, A. J. Willsey, T. W. Yu, R. K. Yuen, D. D. D. Study, Autism Homozygosity Mapping Collaborative for, Uk K. Consortium, E. H. Cook, C. M. Freitag, M. Gill, C. M. Hultman, T. Lehner, A. Palotie, G. D. Schellenberg, P. Sklar, M. W. State, J. S. Sutcliffe, C. A. Walsh, S. W. Scherer, M. E. Zwick, J. C. Barett, D. J. Cutler, K. Roeder, B. Devlin, M. J. Daly, and J. D. Buxbaum. Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature*, 515(7526):209–15, 2014.

[42] M. Dean, M. Carrington, C. Winkler, G. A. Huttley, M. W. Smith, R. Allikmets, J. J. Goedert, S. P. Buchbinder, E. Vittinghoff, E. Gomperts, S. Donfield, D. Vlahov, R. Kaslow, A. Saah, C. Rinaldo, R. Detels, and S. J. O'Brien. Genetic restriction of hiv-1 infection and progression to aids by a deletion allele of the ckr5 structural gene. hemophilia growth and development study, multicenter aids cohort study, multicenter hemophilia cohort study, san francisco city cohort, alive study. *Science*, 273(5283):1856–62, 1996.

[43] M. A. DePristo, E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M. Kernytsky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler, and M. J. Daly. A framework for variation discovery and genotyping using next-generation dna sequencing data. *Nat Genet*, 43(5):491–8, 2011.

[44] Harvard Diabetes Genetics Initiative of Broad Institute of, Lund University Mit, Research Novartis Institutes of BioMedical, R. Saxena, B. F. Voight, V. Lyssenko, N. P. Burtt, P. I. de Bakker, H. Chen, J. J. Roix, S. Kathiresan, J. N. Hirschhorn, M. J. Daly, T. E. Hughes, L. Groop, D. Altshuler, P. Almgren, J. C. Florez, J. Meyer, K. Ardlie, K. Bengtsson Bostrom, B. Isomaa, G. Lettre, U. Lindblad, H. N. Lyon, O. Melander, C. Newton-Cheh, P. Nilsson, M. Orho-Melander, L. Rastam, E. K. Speliotes, M. R. Taskinen, T. Tuomi, C. Guiducci, A. Berglund, J. Carlson, L. Gianniny, R. Hackett, L. Hall, J. Holmkvist, E. Laurila, M. Sjogren, M. Sterner, A. Surti, M. Svensson, M. Svensson, R. Tewhey, B. Blumenstiel, M. Parkin, M. Defelice, R. Barry, W. Brodeur, J. Camarata, N. Chia, M. Fava, J. Gibbons, B. Handsaker, C. Healy, K. Nguyen, C. Gates, C. Sougnez, D. Gage, M. Nizzari, S. B. Gabriel, G. W. Chirn, Q. Ma, H. Parikh, D. Richardson, D. Ricke, and S. Purcell. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science*, 316(5829):1331–6, 2007.

[45] J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, and

B. Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–80, 2012.

[46] L. M. Dong, J. D. Potter, E. White, C. M. Ulrich, L. R. Cardon, and U. Peters. Genetic susceptibility to cancer: the role of polymorphisms in candidate genes. *JAMA*, 299(20):2423–36, 2008.

[47] M. Du, P. L. Auer, S. Jiao, J. Haessler, D. Altshuler, E. Boerwinkle, C. S. Carlson, C. L. Carty, Y. D. Chen, K. Curtis, N. Franceschini, L. Hsu, R. Jackson, L. A. Lange, G. Lettre, K. L. Monda, D. A. Nickerson, A. P. Reiner, S. S. Rich, S. A. Rosse, J. I. Rotter, C. J. Willer, J. G. Wilson, K. North, C. Kooperberg, N. Heard-Costa, and U. Peters. Whole-exome imputation of sequence variants identified two novel alleles associated with adult body height in african americans. *Hum Mol Genet*, 2014.

[48] J. Dupuis, C. Langenberg, I. Prokopenko, R. Saxena, N. Soranzo, A. U. Jackson, E. Wheeler, N. L. Glazer, N. Bouatia-Naji, A. L. Gloyn, C. M. Lindgren, R. Magi, A. P. Morris, J. Randall, T. Johnson, P. Elliott, D. Rybin, G. Thorleifsson, V. Steinthorsdottir, P. Henneman, H. Grallert, A. Dehghan, J. J. Hottenga, C. S. Franklin, P. Navarro, K. Song, A. Goel, J. R. Perry, J. M. Egan, T. Lajunen, N. Grarup, T. Sparso, A. Doney, B. F. Voight, H. M. Stringham, M. Li, S. Kanoni, P. Shrader, C. Cavalcanti-Proenca, M. Kumari, L. Qi, N. J. Timpson, C. Gieger, C. Zabena, G. Rocheleau, E. Ingelsson, P. An, J. O'Connell, J. Luan, A. Elliott, S. A. McCarroll, F. Payne, R. M. Roccasecca, F. Pattou, P. Sethupathy, K. Ardlie, Y. Ariyurek, B. Balkau, P. Barter, J. P. Beilby, Y. Ben-Shlomo, R. Benediktsson, A. J. Bennett, S. Bergmann, M. Bochud, E. Boerwinkle, A. Bonnefond, L. L. Bonnycastle, K. Borch-Johnsen, Y. Bottcher, E. Brunner, S. J. Bumpstead, G. Charpentier, Y. D. Chen, P. Chines, R. Clarke, L. J. Coin, M. N. Cooper, M. Cornelis, G. Crawford, L. Crisponi, I. N. Day, E. J. de Geus, J. Delplanque, C. Dina, M. R. Erdos, A. C. Fedson, A. Fischer-Rosinsky, N. G. Forouhi, C. S. Fox, R. Frants, M. G. Franzosi, P. Galan, M. O. Goodarzi, J. Graessler, C. J. Groves, S. Grundy, R. Gwilliam, U. Gyllensten, S. Hadjadj, et al. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nat Genet*, 42(2):105–16, 2010.

[49] D. Edelman, H. Kalia, M. Delio, M. Alani, K. Krishnamurthy, M. Abd, A. Auton, T. Wang, A. W. Wolkoff, and B. E. Morrow. Genetic analysis of nonalcoholic fatty liver disease within a caribbean-hispanic population. *Mol Genet Genomic Med*, 3(6):558–69, 2015.

[50] D. Enard, L. Cai, C. Gwennap, and D. A. Petrov. Viruses are a dominant driver of protein adaptation in mammals. *Elife*, 5, 2016.

[51] C. Esnault, O. Heidmann, F. Delebecque, M. Dewannieux, D. Ribet, A. J. Hance, T. Heidmann, and O. Schwartz. Apobec3g cytidine deaminase inhibits retrotransposition of endogenous retroviruses. *Nature*, 433(7024):430–3, 2005.

[52] C. Esteban-Jurado, M. Vila-Casadesus, P. Garre, J. J. Lozano, A. Pristoupilova, S. Beltran, J. Munoz, T. Ocana, F. Balaguer, M. Lopez-Ceron, M. Cuatrecasas, S. Franch-Exposito, J. M. Pique, A. Castells, A. Carracedo, C. Ruiz-Ponte, A. Abuli, X. Bessa, M. Andreu, L. Bujanda, T. Caldes, and S. Castellvi-Bel. Whole-exome sequencing identifies rare pathogenic variants in new predisposition genes for familial colorectal cancer. *Genet Med*, 2014.

[53] W.J. Ewens. *Mathematical Population Genetics*. Springer, 2nd edition, 2004.

[54] S. Eyre, J. Bowes, D. Diogo, A. Lee, A. Barton, P. Martin, A. Zhernakova, E. Stahl, S. Viatte, K. McAllister, C. I. Amos, L. Padyukov, R. E. Toes, T. W. Huizinga, C. Wijmenga, G. Trynka, L. Franke, H. J. Westra, L. Alfredsson, X. Hu, C. Sandor, P. I. de Bakker, S. Davila, C. C. Khor, K. K. Heng, R. Andrews, S. Edkins, S. E. Hunt, C. Langford, D. Symmons, Genetics Biologics in Rheumatoid Arthritis, Syndicate Genomics Study, Consortium Wellcome Trust Case Control, P. Concannon, S. Onengut-Gumuscu, S. S. Rich, P. Deloukas, M. A. Gonzalez-Gay, L. Rodriguez-Rodriguez, L. Arlsetig, J. Martin, S. Rantapaa-Dahlqvist, R. M. Plenge, S. Raychaudhuri, L. Klareskog, P. K. Gregersen, and J. Worthington. High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. *Nat Genet*, 44(12):1336–40, 2012.

[55] Adam Eyre-Walker. Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. *Proceedings of the National Academy of Sciences*, 107(suppl 1):1752–1756, 2010.

[56] M. F. Feitosa, M. K. Wojczynski, K. E. North, Q. Zhang, M. A. Province, J. J. Carr, and I. B. Borecki. The erlin1-chuk-cwf19l1 gene cluster influences liver fat deposition and hepatic inflammation in the nhlbi family heart study. *Atherosclerosis*, 228(1):175–80, 2013.

[57] J. Fellay, K. V. Shianna, D. Ge, S. Colombo, B. Ledergerber, M. Weale, K. Zhang, C. Gumbs, A. Castagna, A. Cossarizza, A. Cozzi-Lepri, A. De Luca, P. Easterbrook, P. Francioli, S. Mallal, J. Martinez-Picado, J. M. Miro, N. Obel, J. P. Smith, J. Wyniger, P. Descombes, S. E. Antonarakis, N. L.

Letvin, A. J. McMichael, B. F. Haynes, A. Telenti, and D. B. Goldstein. A whole-genome association study of major determinants for host control of hiv-1. *Science*, 317(5840):944–7, 2007.

[58] J. Fellay, K. V. Shianna, A. Telenti, and D. B. Goldstein. Host genetics and hiv-1: the final phase? *PLoS Pathog*, 6(10):e1001033, 2010.

[59] W. Feller. *An Introduction to Probability Theory and its Applications*. Wiley, 1968.

[60] M. Fernandez, W. Raskind, J. Wolff, M. Matsushita, E. Yuen, W. Graf, H. Lipe, and T. Bird. Familial dyskinesia and facial myokymia (fdfm): a novel movement disorder. *Ann Neurol*, 49(4):486–92, 2001.

[61] M. A. Ferreira, M. C. O'Donovan, Y. A. Meng, I. R. Jones, D. M. Ruderfer, L. Jones, J. Fan, G. Kirov, R. H. Perlis, E. K. Green, J. W. Smoller, D. Grozeva, J. Stone, I. Nikolov, K. Chambert, M. L. Hamshere, V. L. Nimgaonkar, V. Moskvina, M. E. Thase, S. Caesar, G. S. Sachs, J. Franklin, K. Gordon-Smith, K. G. Ardlie, S. B. Gabriel, C. Fraser, B. Blumenstiel, M. Defelice, G. Breen, M. Gill, D. W. Morris, A. Elkin, W. J. Muir, K. A. McGhee, R. Williamson, D. J. MacIntyre, A. W. MacLean, C. D. St, M. Robinson, M. Van Beck, A. C. Pereira, R. Kandaswamy, A. McQuillin, D. A. Collier, N. J. Bass, A. H. Young, J. Lawrence, I. N. Ferrier, A. Anjorin, A. Farmer, D. Curtis, E. M. Scolnick, P. McGuffin, M. J. Daly, A. P. Corvin, P. A. Holmans, D. H. Blackwood, H. M. Gurling, M. J. Owen, S. M. Purcell, P. Sklar, N. Craddock, and Consortium Wellcome Trust Case Control. Collaborative genome-wide association analysis supports a role for ank3 and cacna1c in bipolar disorder. *Nat Genet*, 40(9):1056–8, 2008.

[62] M. A. Ferreira and S. M. Purcell. A multivariate test of association. *Bioinformatics*, 25(1):132–3, 2009.

[63] J. C. Florez, J. Hirschhorn, and D. Altshuler. The inherited basis of diabetes mellitus: implications for the genetic analysis of complex traits. *Annu Rev Genomics Hum Genet*, 4:257–91, 2003.

[64] K. A. Frazer, S. S. Murray, N. J. Schork, and E. J. Topol. Human genetic variation and its contribution to complex traits. *Nat Rev Genet*, 10(4):241–51, 2009.

[65] R. M. Freathy, D. O. Mook-Kanamori, U. Sovio, I. Prokopenko, N. J. Timpson, D. J. Berry, N. M. Warrington, E. Widen, J. J. Hottenga, M. Kaakinen, L. A. Lange, J. P. Bradfield, M. Kerkhof, J. A. Marsh, R. Magi, C. M. Chen, H. N. Lyon, M. Kirin, L. S. Adair, Y. S. Aulchenko, A. J. Bennett, J. B. Borja, N. Bouatia-Naji, P. Charoen, L. J. Coin, D. L. Cousminer, E. J. de Geus, P. Deloukas, P. Elliott, D. M. Evans, P. Froguel, ANthropometric Traits Consortium Genetic Investigation of, B. Glaser, C. J. Groves, A. L. Hartikainen, N. Hassanali, J. N. Hirschhorn, A. Hofman, J. M. Holly, E. Hypponen, S. Kanoni, B. A. Knight, J. Laitinen, C. M. Lindgren, Glucose Meta-Analyses of, Consortium Insulin-related traits, W. L. McArdle, P. F. O'Reilly, C. E. Pennell, D. S. Postma, A. Pouta, A. Ramasamy, N. W. Rayner, S. M. Ring, F. Rivadeneira, B. M. Shields, D. P. Strachan, I. Surakka, A. Taanila, C. Tiesler, A. G. Uitterlinden, C. M. van Duijn, Consortium Wellcome Trust Case Control, A. H. Wijga, G. Willemsen, H. Zhang, J. Zhao, J. F. Wilson, E. A. Steegers, A. T. Hattersley, J. G. Eriksson, L. Peltonen, K. L. Mohlke, S. F. Grant, H. Hakonarson, G. H. Koppelman, G. V. Dedoussis, J. Heinrich, M. W. Gillman, L. J. Palmer, T. M. Frayling, D. I. Boomsma, G. Davey Smith, C. Power, V. W. Jaddoe, M. R. Jarvelin, Consortium Early Growth Genetics, and M. I. McCarthy. Variants in adcy5 and near ccnl1 are associated with fetal growth and birth weight. *Nat Genet*, 42(5):430–5, 2010.

[66] H. Freeman and R. D. Cox. Type-2 diabetes: a cocktail of genetic discovery. *Hum Mol Genet*, 15 Spec No 2:R202–9, 2006.

[67] W. Fu, T. D. O'Connor, and J. M. Akey. Genetic architecture of quantitative traits and complex diseases. *Curr Opin Genet Dev*, 23(6):678–83, 2013.

[68] Wenqing Fu, Timothy D OConnor, Goo Jun, Hyun Min Kang, Goncalo Abecasis, Suzanne M Leal, Stacey Gabriel, David Altshuler, Jay Shendure, Deborah A Nickerson, et al. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature*, 2012.

[69] M. J. Fullwood, M. H. Liu, Y. F. Pan, J. Liu, H. Xu, Y. B. Mohamed, Y. L. Orlov, S. Velkov, A. Ho, P. H. Mei, E. G. Chew, P. Y. Huang, W. J. Welboren, Y. Han, H. S. Ooi, P. N. Ariyaratne, V. B. Vega, Y. Luo, P. Y. Tan, P. Y. Choy, K. D. Wansa, B. Zhao, K. S. Lim, S. C. Leow, J. S. Yow, R. Joseph, H. Li, K. V. Desai, J. S. Thomsen, Y. K. Lee, R. K. Karuturi, T. Herve, G. Bourque, H. G. Stunnenberg, X. Ruan, V. Cacheux-Rataboul, W. K. Sung, E. T. Liu, C. L. Wei, E. Cheung, and Y. Ruan. An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature*, 462(7269):58–64, 2009.

[70] S. B. Gabriel, S. F. Schaffner, H. Nguyen, J. M. Moore, J. Roy, B. Blumenstiel, J. Higgins, M. DeFelice, A. Lochner, M. Faggart, S. N. Liu-Cordero, C. Rotimi, A. Adeyemo, R. Cooper, R. Ward, E. S. Lander, M. J. Daly, and D. Altshuler. The structure of haplotype blocks in the human genome. *Science*, 296(5576):2225–9, 2002.

[71] D. J. Gaffney. Global properties and functional complexity of human gene regulatory variation. *PLoS Genet*, 9(5):e1003501, 2013.

[72] T. E. Galesloot, K. van Steen, L. A. Kiemeney, L. L. Janss, and S. H. Vermeulen. A comparison of multivariate genome-wide association methods. *PLoS One*, 9(4):e95923, 2014.

[73] Consortium Genomes Project, A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison, H. M. Kang, J. O. Korbel, J. L. Marchini, S. McCarthy, G. A. McVean, and G. R. Abecasis. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

[74] G. Gibson. Rare and common variants: twenty arguments. *Nature Reviews Genetics*, 13(2):135–145, 2012.

[75] C. Gilissen, J. Y. Hehir-Kwa, D. T. Thung, M. van de Vorst, B. W. van Bon, M. H. Willemsen, M. Kwint, I. M. Janssen, A. Hoischen, A. Schenck, R. Leach, R. Klein, R. Tearle, T. Bo, R. Pfundt, H. G. Yntema, B. B. de Vries, T. Kleefstra, H. G. Brunner, L. E. Vissers, and J. A. Veltman. Genome sequencing identifies major causes of severe intellectual disability. *Nature*, 511(7509):344–7, 2014.

[76] J.H. Gillespie. *Population Genetics: A Concise Guide*. Johns Hopkins University Press, 2nd edition, 2004.

[77] A. Gilly, K. Kuchenbaecker, L. Southam, D. Suveges, R. Moore, G. Melloni, K. Hatzikotoulas, A. Farmaki, G. Ritchie, J. Schwartzentruber, P. Danecek, B. Kilian, M. Pollard, X. Ge, H. Elding, W. Astle, T. Jiang, A. Butterworth, N. Soranzo, E. Tsafantakis, M. Karaleftheri, G. Dedoussis, and E. Zeggini. Very low depth whole genome sequencing in complex trait association studies. *bioRxiv*, 2017.

[78] D. Goldman. Candidate genes in alcoholism. *Clin Neurosci*, 3(3):174–81, 1995.

[79] D. B. Goldstein, A. Allen, J. Keebler, E. H. Margulies, S. Petrou, S. Petrovski, and S. Sunyaev. Sequencing studies in human genetics: design and interpretation. *Nat Rev Genet*, 14(7):460–70, 2013.

[80] O. Gottesman, H. Kuivaniemi, G. Tromp, W. A. Faucett, R. Li, T. A. Manolio, S. C. Sanderson, J. Kannry, R. Zinberg, M. A. Basford, M. Brilliant, D. J. Carey, R. L. Chisholm, C. G. Chute, J. J. Connolly, D. Crosslin, J. C. Denny, C. J. Gallego, J. L. Haines, H. Hakonarson, J. Harley, G. P. Jarvik, I. Kohane, I. J. Kullo, E. B. Larson, C. McCarty, M. D. Ritchie, D. M. Roden, M. E. Smith, E. P. Bottinger, M. S. Williams, and Merge Network e. The electronic medical records and genomics (emerge) network: past, present, and future. *Genet Med*, 15(10):761–71, 2013.

[81] S. Gravel, B. M. Henn, R. N. Gutenkunst, A. R. Indap, G. T. Marth, A. G. Clark, F. Yu, R. A. Gibbs, and C. D. Bustamante. Demographic history and rare allele sharing among human populations. *Proc Natl Acad Sci U S A*, 108(29):11983–8, 2011.

[82] R. E. Green, J. Krause, A. W. Briggs, T. Maricic, U. Stenzel, M. Kircher, N. Patterson, H. Li, W. Zhai, M. H. Fritz, N. F. Hansen, E. Y. Durand, A. S. Malaspinas, J. D. Jensen, T. Marques-Bonet, C. Alkan, K. Prufer, M. Meyer, H. A. Burbano, J. M. Good, R. Schultz, A. Aximu-Petri, A. Butthof, B. Hober, B. Hoffner, M. Siegemund, A. Weihmann, C. Nusbaum, E. S. Lander, C. Russ, N. Novod, J. Affourtit, M. Egholm, C. Verna, P. Rudan, D. Brajkovic, Z. Kucan, I. Gusic, V. B. Doronichev, L. V. Golovanova, C. Lalueza-Fox, M. de la Rasilla, J. Fortea, A. Rosas, R. W. Schmitz, P. L. F. Johnson, E. E. Eichler, D. Falush, E. Birney, J. C. Mullikin, M. Slatkin, R. Nielsen, J. Kelso, M. Lachmann, D. Reich, and S. Paabo. A draft sequence of the neandertal genome. *Science*, 328(5979):710–722, 2010.

[83] J. F. Gusella and M. E. MacDonald. Huntington's disease: the case for genetic modifiers. *Genome Med*, 1(8):80, 2009.

[84] Ryan N Gutenkunst, Ryan D Hernandez, Scott H Williamson, and Carlos D Bustamante. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics*, 5(10):e1000695, 2009.

[85] R. E. Haaland, J. A. Johnson, and J. Tang. Recent advances in research of hiv infection: implications of viral and host genetics on treatment and prevention. *Public Health Genomics*, 16(1-2):31–6, 2013.

[86] A. T. Hattersley and J. E. Tooke. The fetal insulin hypothesis: an alternative explanation of the association of low birthweight with diabetes and vascular disease. *Lancet*, 353(9166):1789–92, 1999.

[87] I. M. Heid, A. U. Jackson, J. C. Randall, T. W. Winkler, L. Qi, V. Steinthorsdottir, G. Thorleifsson, M. C. Zillikens, E. K. Speliotes, R. Magi, T. Workalemahu, C. C. White, N. Bouatia-Naji, T. B. Harris, S. I. Berndt, E. Ingelsson, C. J. Willer, M. N. Weedon, J. Luan, S. Vedantam, T. Esko, T. O. Kilpelainen, Z. Kutalik, S. Li, K. L. Monda, A. L. Dixon, C. C. Holmes, L. M. Kaplan, L. Liang, J. L. Min, M. F. Moffatt, C. Molony, G. Nicholson, E. E. Schadt, K. T. Zondervan, M. F. Feitosa, T. Ferreira, H. Lango Allen, R. J. Weyant, E. Wheeler, A. R. Wood, Magic, K. Estrada, M. E. Goddard, G. Lettre, M. Mangino, D. R. Nyholt, S. Purcell, A. V. Smith, P. M. Visscher, J. Yang, S. A. McCarroll, J. Nemesh, B. F. Voight, D. Absher, N. Amin, T. Aspelund, L. Coin, N. L. Glazer, C. Hayward, N. L. Heard-Costa, J. J. Hottenga, A. Johansson, T. Johnson, M. Kaakinen, K. Kapur, S. Ketkar, J. W. Knowles, P. Kraft, A. T. Kraja, C. Lamina, M. F. Leitzmann, B. McKnight, A. P. Morris, K. K. Ong, J. R. Perry, M. J. Peters, O. Polasek, I. Prokopenko, N. W. Rayner, S. Ripatti, F. Rivadeneira, N. R. Robertson, S. Sanna, U. Sovio, I. Surakka, A. Teumer, S. van Wingerden, V. Vitart, J. H. Zhao, C. Cavalcanti-Proenca, P. S. Chines, E. Fisher, J. R. Kulzer, C. Lecoeur, N. Narisu, C. Sandholt, L. J. Scott, K. Silander, K. Stark, et al. Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. *Nat Genet*, 42(11):949–60, 2010.

[88] D. P. Hibar, J. L. Stein, M. E. Renteria, A. Arias-Vasquez, S. Desrivieres, N. Jahanshad, R. Toro, K. Wittfeld, L. Abramovic, M. Andersson, B. S. Aribisala, N. J. Armstrong, M. Bernard, M. M. Bohlken, M. P. Boks, J. Bralten, A. A. Brown, M. M. Chakravarty, Q. Chen, C. R. Ching, G. Cuellar-Partida, A. den Braber, S. Giddaluru, A. L. Goldman, O. Grimm, T. Guadalupe, J. Hass, G. Woldehawariat, A. J. Holmes, M. Hoogman, D. Janowitz, T. Jia, S. Kim, M. Klein, B. Kraemer, P. H. Lee, L. M. Olde Loohuis, M. Luciano, C. Macare, K. A. Mather, M. Mattheisen, Y. Milaneschi, K. Nho, M. Papmeyer, A. Ramasamy, S. L. Risacher, R. Roiz-Santianez, E. J. Rose, A. Salami, P. G. Samann, L. Schmaal, A. J. Schork, J. Shin, L. T. Strike, A. Teumer, M. M. van Donkelaar, K. R. van Eijk, R. K. Walters, L. T. Westlye, C. D. Whelan, A. M. Winkler, M. P. Zwiers, S. Alhusaini, L. Athanasiu, S. Ehrlich, M. M. Hakobjan, C. B. Hartberg, U. K. Haukvik, A. J. Heister, D. Hoehn, D. Kasperaviciute, D. C. Liewald, L. M. Lopez, R. R. Makkinje, M. Matarin, M. A. Naber, D. R.

McKay, M. Needham, A. C. Nugent, B. Putz, N. A. Royle, L. Shen, E. Sprooten, D. Trabzuni, S. S. van der Marel, K. J. van Hulzen, E. Walton, C. Wolf, L. Almasy, D. Ames, S. Arepalli, A. A. Assareh, M. E. Bastin, H. Brodaty, K. B. Bulayeva, M. A. Carless, S. Cichon, A. Corvin, J. E. Curran, M. Czisch, et al. Common genetic variants influence human subcortical brain structures. *Nature*, 520(7546):224–9, 2015.

[89] L. A. Hindorff, P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins, and T. A. Manolio. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A*, 106(23):9362–7, 2009.

[90] J. N. Hirschhorn and M. J. Daly. Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet*, 6(2):95–108, 2005.

[91] D. J. Hodson, R. K. Mitchell, L. Marselli, T. J. Pullen, S. Gimeno Brias, F. Semplici, K. L. Everett, D. M. Cooper, M. Bugliani, P. Marchetti, V. Lavallard, D. Bosco, L. Piemonti, P. R. Johnson, S. J. Hughes, D. Li, W. H. Li, A. M. Shapiro, and G. A. Rutter. Adcy5 couples glucose to insulin secretion in human islets. *Diabetes*, 63(9):3009–21, 2014.

[92] P. D. Hsu, E. S. Lander, and F. Zhang. Development and applications of crispr-cas9 for genome engineering. *Cell*, 157(6):1262–78, 2014.

[93] R.R. Hudson. Gene genealogies and the coalescent process. *Oxf. Surv. Evol. Biol.*, 7:144, 1990.

[94] H. B. Hutcheson, J. A. Lautenberger, G. W. Nelson, J. U. Pontius, B. D. Kessing, C. A. Winkler, M. W. Smith, R. Johnson, R. Stephens, J. Phair, J. J. Goedert, S. Donfield, and S. J. O'Brien. Detecting aids restriction genes: from candidate genes to genome-wide association discovery. *Vaccine*, 26(24):2951–65, 2008.

[95] Studies International Consortium for Blood Pressure Genome-Wide Association, G. B. Ehret, P. B. Munroe, K. M. Rice, M. Bochud, A. D. Johnson, D. I. Chasman, A. V. Smith, M. D. Tobin, G. C. Verwoert, S. J. Hwang, V. Pihur, P. Vollenweider, P. F. O'Reilly, N. Amin, J. L. Bragg-Gresham, A. Teumer, N. L. Glazer, L. Launer, J. H. Zhao, Y. Aulchenko, S. Heath, S. Sober, A. Parsa, J. Luan, P. Arora, A. Dehghan, F. Zhang, G. Lucas, A. A. Hicks, A. U. Jackson, J. F. Peden, T. Tanaka, S. H. Wild, I. Rudan, W. Igl, Y. Milaneschi, A. N. Parker, C. Fava, J. C. Chambers, E. R. Fox,

M. Kumari, M. J. Go, P. van der Harst, W. H. Kao, M. Sjogren, D. G. Vinay, M. Alexander, Y. Tabara, S. Shaw-Hawkins, P. H. Whincup, Y. Liu, G. Shi, J. Kuusisto, B. Tayo, M. Seielstad, X. Sim, K. D. Nguyen, T. Lehtimaki, G. Matullo, Y. Wu, T. R. Gaunt, N. C. Onland-Moret, M. N. Cooper, C. G. Platou, E. Org, R. Hardy, S. Dahgam, J. Palmen, V. Vitart, P. S. Braund, T. Kuznetsova, C. S. Uiterwaal, A. Adeyemo, W. Palmas, H. Campbell, B. Ludwig, M. Tomaszewski, I. Tzoulaki, N. D. Palmer, C. ARDIoGRAM consortium, C. KDGen Consortium, Consortium KidneyGen, consortium EchoGen, Charge-Hf consortium, T. Aspelund, M. Garcia, Y. P. Chang, J. R. O'Connell, N. I. Steinle, D. E. Grobbee, D. E. Arking, S. L. Kardia, A. C. Morrison, D. Hernandez, S. Najjar, W. L. McArdle, D. Hadley, M. J. Brown, J. M. Connell, et al. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature*, 478(7367):103–9, 2011.

[96] Consortium International HapMap. The international hapmap project. *Nature*, 426(6968):789–96, 2003.

[97] Consortium International HapMap. A haplotype map of the human genome. *Nature*, 437(7063):1299–320, 2005.

[98] Consortium International HapMap, D. M. Altshuler, R. A. Gibbs, L. Peltonen, D. M. Altshuler, R. A. Gibbs, L. Peltonen, E. Dermitzakis, S. F. Schaffner, F. Yu, L. Peltonen, E. Dermitzakis, P. E. Bonnen, D. M. Altshuler, R. A. Gibbs, P. I. de Bakker, P. Deloukas, S. B. Gabriel, R. Gwilliam, S. Hunt, M. Inouye, X. Jia, A. Palotie, M. Parkin, P. Whittaker, F. Yu, K. Chang, A. Hawes, L. R. Lewis, Y. Ren, D. Wheeler, R. A. Gibbs, D. M. Muzny, C. Barnes, K. Darvishi, M. Hurles, J. M. Korn, K. Kristiansson, C. Lee, S. A. McCarrol, J. Nemesh, E. Dermitzakis, A. Keinan, S. B. Montgomery, S. Pollack, A. L. Price, N. Soranzo, P. E. Bonnen, R. A. Gibbs, C. Gonzaga-Jauregui, A. Keinan, A. L. Price, F. Yu, V. Anttila, W. Brodeur, M. J. Daly, S. Leslie, G. McVean, L. Moutsianas, H. Nguyen, S. F. Schaffner, Q. Zhang, M. J. Ghori, R. McGinnis, W. McLaren, S. Pollack, A. L. Price, S. F. Schaffner, F. Takeuchi, S. R. Grossman, I. Shlyakhter, E. B. Hostetter, P. C. Sabeti, C. A. Adebamowo, M. W. Foster, D. R. Gordon, J. Licinio, M. C. Manca, P. A. Marshall, I. Matsuda, D. Ngare, V. O. Wang, D. Reddy, C. N. Rotimi, C. D. Royal, R. R. Sharp, C. Zeng, L. D. Brooks, and J. E. McEwen. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52–8, 2010.

[99] Consortium International HapMap, K. A. Frazer, D. G. Ballinger, D. R. Cox,

D. A. Hinds, L. L. Stuve, R. A. Gibbs, J. W. Belmont, A. Boudreau, P. Hardenbol, S. M. Leal, S. Pasternak, D. A. Wheeler, T. D. Willis, F. Yu, H. Yang, C. Zeng, Y. Gao, H. Hu, W. Hu, C. Li, W. Lin, S. Liu, H. Pan, X. Tang, J. Wang, W. Wang, J. Yu, B. Zhang, Q. Zhang, H. Zhao, H. Zhao, J. Zhou, S. B. Gabriel, R. Barry, B. Blumenstiel, A. Camargo, M. Defelice, M. Faggart, M. Goyette, S. Gupta, J. Moore, H. Nguyen, R. C. Onofrio, M. Parkin, J. Roy, E. Stahl, E. Winchester, L. Ziaugra, D. Altshuler, Y. Shen, Z. Yao, W. Huang, X. Chu, Y. He, L. Jin, Y. Liu, Y. Shen, W. Sun, H. Wang, Y. Wang, Y. Wang, X. Xiong, L. Xu, M. M. Waye, S. K. Tsui, H. Xue, J. T. Wong, L. M. Galver, J. B. Fan, K. Gunderson, S. S. Murray, A. R. Oliphant, M. S. Chee, A. Montpetit, F. Chagnon, V. Ferretti, M. Leboeuf, J. F. Olivier, M. S. Phillips, S. Roumy, C. Sallee, A. Verner, T. J. Hudson, P. Y. Kwok, D. Cai, D. C. Koboldt, R. D. Miller, L. Pawlikowska, P. Taillon-Miller, M. Xiao, L. C. Tsui, W. Mak, Y. Q. Song, P. K. Tam, Y. Nakamura, T. Kawaguchi, T. Kitamoto, T. Morizono, A. Nagashima, et al. A second generation human haplotype map of over 3.1 million snps. *Nature*, 449(7164):851–61, 2007.

[100] Consortium International Schizophrenia, S. M. Purcell, N. R. Wray, J. L. Stone, P. M. Visscher, M. C. O'Donovan, P. F. Sullivan, and P. Sklar. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, 460(7256):748–52, 2009.

[101] J. P. Ioannidis, E. E. Ntzani, T. A. Trikalinos, and D. G. Contopoulos-Ioannidis. Replication validity of genetic association studies. *Nat Genet*, 29(3):306–9, 2001.

[102] T. Iwamoto, S. Okumura, K. Iwatsubo, J. Kawabe, K. Ohtsu, I. Sakai, Y. Hashimoto, A. Izumitani, K. Sango, K. Ajiki, Y. Toya, S. Umemura, Y. Goshima, N. Arai, S. F. Vatner, and Y. Ishikawa. Motor dysfunction in type 5 adenylyl cyclase-null mice. *J Biol Chem*, 278(19):16936–40, 2003.

[103] A. L. Jackson and P. S. Linsley. Recognizing and avoiding sirna off-target effects for target identification and therapeutic application. *Nat Rev Drug Discov*, 9(1):57–67, 2010.

[104] S. Jager, P. Cimermancic, N. Gulbahce, J. R. Johnson, K. E. McGovern, S. C. Clarke, M. Shales, G. Mercenne, L. Pache, K. Li, H. Hernandez, G. M. Jang, S. L. Roth, E. Akiva, J. Marlett, M. Stephens, I. D'Orso, J. Fernandes, M. Fahey, C. Mahon, A. J. O'Donoghue, A. Todorovic, J. H. Morris, D. A. Maltby, T. Alber, G. Cagney, F. D. Bushman, J. A. Young, S. K. Chanda,

W. I. Sundquist, T. Kortemme, R. D. Hernandez, C. S. Craik, A. Burlingame, A. Sali, A. D. Frankel, and N. J. Krogan. Global landscape of hiv-human protein complexes. *Nature*, 481(7381):365–70, 2012.

[105] C. Jiang and Z. B. Zeng. Multiple trait analysis of genetic mapping for quantitative trait loci. *Genetics*, 140(3):1111–27, 1995.

[106] M. Jinek, K. Chylinski, I. Fonfara, M. Hauer, J. A. Doudna, and E. Charpentier. A programmable dual-rna-guided dna endonuclease in adaptive bacterial immunity. *Science*, 337(6096):816–21, 2012.

[107] G. C. Johnson, L. Esposito, B. J. Barratt, A. N. Smith, J. Heward, G. Di Genova, H. Ueda, H. J. Cordell, I. A. Eaves, F. Dudbridge, R. C. Twells, F. Payne, W. Hughes, S. Nutland, H. Stevens, P. Carr, E. Tuomilehto-Wolf, J. Tuomilehto, S. C. Gough, D. G. Clayton, and J. A. Todd. Haplotype tagging for the identification of common disease genes. *Nat Genet*, 29(2):233–7, 2001.

[108] Toby Johnson and Nick Barton. Theoretical models of selection and mutation on quantitative traits. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1459):1411–1425, 2005.

[109] I. Jones and N. Craddock. Candidate gene studies of bipolar disorder. *Ann Med*, 33(4):248–56, 2001.

[110] R. A. Kaslow, D. G. Ostrow, R. Detels, J. P. Phair, B. F. Polk, and Jr. Rinaldo, C. R. The multicenter aids cohort study: rationale, organization, and selected characteristics of the participants. *Am J Epidemiol*, 126(2):310–8, 1987.

[111] J. Kawalkowska, A. M. Quirke, F. Ghari, S. Davis, V. Subramanian, P. R. Thompson, R. O. Williams, R. Fischer, N. B. La Thangue, and P. J. Venables. Abrogation of collagen-induced arthritis by a peptidyl arginine deiminase inhibitor is associated with modulation of t cell-mediated immune responses. *Sci Rep*, 6:26430, 2016.

[112] A. Keinan and A. G. Clark. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science*, 336(6082):740–743, 2012.

[113] Alon Keinan, James C Mullikin, Nick Patterson, and David Reich. Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. *Nature Genetics*, 39(10):1251–1255, 2007.

[114] M. C. King and A. C. Wilson. Evolution at two levels in humans and chimpanzees. *Science*, 188(4184):107–16, 1975.

[115] L. Klei, D. Luca, B. Devlin, and K. Roeder. Pleiotropy and principal components of heritability combine to increase power for association analysis. *Genet Epidemiol*, 32(1):9–19, 2008.

[116] A. Kong, M. L. Frigge, G. Masson, S. Besenbacher, P. Sulem, G. Magnusson, S. A. Gudjonsson, A. Sigurdsson, A. Jonasdottir, A. Jonasdottir, W. S. Wong, G. Sigurdsson, G. B. Walters, S. Steinberg, H. Helgason, G. Thorleifsson, D. F. Gudbjartsson, A. Helgason, O. T. Magnusson, U. Thorsteinsdottir, and K. Stefansson. Rate of de novo mutations and the importance of father's age to disease risk. *Nature*, 488(7412):471–5, 2012.

[117] R. Konig, Y. Zhou, D. Elleder, T. L. Diamond, G. M. Bonamy, J. T. Irelan, C. Y. Chiang, B. P. Tu, P. D. De Jesus, C. E. Lilley, S. Seidel, A. M. Opaluch, J. S. Caldwell, M. D. Weitzman, K. L. Kuhen, S. Bandyopadhyay, T. Ideker, A. P. Orth, L. J. Miraglia, F. D. Bushman, J. A. Young, and S. K. Chanda. Global analysis of host-pathogen interactions that regulate early-stage hiv-1 replication. *Cell*, 135(1):49–60, 2008.

[118] M. Korb, A. G. Rust, V. Thorsson, C. Battail, B. Li, D. Hwang, K. A. Kennedy, J. C. Roach, C. M. Rosenberger, M. Gilchrist, D. Zak, C. Johnson, B. Marzolf, A. Aderem, I. Shmulevich, and H. Bolouri. The innate immune database (iidb). *BMC Immunol*, 9:7, 2008.

[119] A. Kottgen, C. Pattaro, C. A. Boger, C. Fuchsberger, M. Olden, N. L. Glazer, A. Parsa, X. Gao, Q. Yang, A. V. Smith, J. R. O'Connell, M. Li, H. Schmidt, T. Tanaka, A. Isaacs, S. Ketkar, S. J. Hwang, A. D. Johnson, A. Dehghan, A. Teumer, G. Pare, E. J. Atkinson, T. Zeller, K. Lohman, M. C. Cornelis, N. M. Probst-Hensch, F. Kronenberg, A. Tonjes, C. Hayward, T. Aspelund, G. Eiriksdottir, L. J. Launer, T. B. Harris, E. Rampersaud, B. D. Mitchell, D. E. Arking, E. Boerwinkle, M. Struchalin, M. Cavalieri, A. Singleton, F. Giallauria, J. Metter, I. H. de Boer, T. Haritunians, T. Lumley, D. Siscovick, B. M. Psaty, M. C. Zillikens, B. A. Oostra, M. Feitosa, M. Province, M. de Andrade, S. T. Turner, A. Schillert, A. Ziegler, P. S. Wild, R. B. Schnabel, S. Wilde, T. F. Munzel, T. S. Leak, T. Illig, N. Klopp, C. Meisinger, H. E. Wichmann, W. Koenig, L. Zgaga, T. Zemunik, I. Kolcic, C. Minelli, F. B. Hu, A. Johansson, W. Igl, G. Zaboli, S. H. Wild, A. F. Wright, H. Campbell, D. Ellinghaus, S. Schreiber, Y. S. Aulchenko, J. F. Felix, F. Rivadeneira, A. G. Uitterlinden,

A. Hofman, M. Imboden, D. Nitsch, A. Brandstatter, B. Kollerits, L. Kedenko, R. Magi, M. Stumvoll, P. Kovacs, M. Boban, S. Campbell, K. Endlich, H. Volzke, H. K. Kroemer, M. Nauck, U. Volker, O. Polasek, V. Vitart, et al. New loci associated with kidney function and chronic kidney disease. *Nat Genet*, 42(5):376–84, 2010.

[120] P. Kumar, S. Henikoff, and P. C. Ng. Predicting the effects of coding non-synonymous variants on protein function using the sift algorithm. *Nat Protoc*, 4(7):1073–81, 2009.

[121] Prateek Kumar, Steven Henikoff, and Pauline C Ng. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols*, 4(7):1073–1081, 2009.

[122] T. A. Kunkel. Rapid and efficient site-specific mutagenesis without phenotypic selection. *Proc Natl Acad Sci U S A*, 82(2):488–92, 1985.

[123] J. M. Kwon and A. M. Goate. The candidate gene approach. *Alcohol Res Health*, 24(3):164–8, 2000.

[124] E. S. Lander. The new genomics: global views of biology. *Science*, 274(5287):536–9, 1996.

[125] E. S. Lander. Initial impact of the sequencing of the human genome. *Nature*, 470(7333):187–97, 2011.

[126] E. S. Lander and N. J. Schork. Genetic dissection of complex traits. *Science*, 265(5181):2037–48, 1994.

[127] L. A. Lange, Y. Hu, H. Zhang, C. Xue, E. M. Schmidt, Z. Z. Tang, C. Bizon, E. M. Lange, J. D. Smith, E. H. Turner, G. Jun, H. M. Kang, G. Peloso, P. Auer, K. P. Li, J. Flannick, J. Zhang, C. Fuchsberger, K. Gaulton, C. Lindgren, A. Locke, A. Manning, X. Sim, M. A. Rivas, O. L. Holmen, O. Gottesman, Y. Lu, D. Ruderfer, E. A. Stahl, Q. Duan, Y. Li, P. Durda, S. Jiao, A. Isaacs, A. Hofman, J. C. Bis, A. Correa, M. E. Griswold, J. Jakobsdottir, A. V. Smith, P. J. Schreiner, M. F. Feitosa, Q. Zhang, J. E. Huffman, J. Crosby, C. L. Wassel, R. Do, N. Franceschini, L. W. Martin, J. G. Robinson, T. L. Assimes, D. R. Crosslin, E. A. Rosenthal, M. Tsai, M. J. Rieder, D. N. Farlow, A. R. Folsom, T. Lumley, E. R. Fox, C. S. Carlson, U. Peters, R. D. Jackson, C. M. van Duijn, A. G. Uitterlinden, D. Levy, J. I. Rotter, H. A. Taylor, Jr. Gudnason, V., D. S. Siscovick, M. Fornage, I. B. Borecki, C. Hayward, I. Rudan, Y. E. Chen, E. P.

Bottinger, R. J. Loos, P. Saetrom, K. Hveem, M. Boehnke, L. Groop, M. McCarthy, T. Meitinger, C. M. Ballantyne, S. B. Gabriel, C. J. O'Donnell, W. S. Post, K. E. North, A. P. Reiner, E. Boerwinkle, B. M. Psaty, D. Altshuler, S. Kathiresan, D. Y. Lin, G. P. Jarvik, L. A. Cupples, C. Kooperberg, J. G. Wilson, D. A. Nickerson, G. R. Abecasis, S. S. Rich, et al. Whole-exome sequencing identifies rare and low-frequency coding variants associated with ldl cholesterol. *Am J Hum Genet*, 94(2):233–45, 2014.

[128] H. Lango Allen, K. Estrada, G. Lettre, S. I. Berndt, M. N. Weedon, F. Rivadeneira, C. J. Willer, A. U. Jackson, S. Vedantam, S. Raychaudhuri, T. Ferreira, A. R. Wood, R. J. Weyant, A. V. Segre, E. K. Speliotes, E. Wheeler, N. Soranzo, J. H. Park, J. Yang, D. Gudbjartsson, N. L. Heard-Costa, J. C. Randall, L. Qi, A. Vernon Smith, R. Magi, T. Pastinen, L. Liang, I. M. Heid, J. Luan, G. Thorleifsson, T. W. Winkler, M. E. Goddard, K. Sin Lo, C. Palmer, T. Workalemahu, Y. S. Aulchenko, A. Johansson, M. C. Zillikens, M. F. Feitosa, T. Esko, T. Johnson, S. Ketkar, P. Kraft, M. Mangino, I. Prokopenko, D. Absher, E. Albrecht, F. Ernst, N. L. Glazer, C. Hayward, J. J. Hottenga, K. B. Jacobs, J. W. Knowles, Z. Kutalik, K. L. Monda, O. Polasek, M. Preuss, N. W. Rayner, N. R. Robertson, V. Steinthorsdottir, J. P. Tyrer, B. F. Voight, F. Wiklund, J. Xu, J. H. Zhao, D. R. Nyholt, N. Pellikka, M. Perola, J. R. Perry, I. Surakka, M. L. Tammesoo, E. L. Altmaier, N. Amin, T. Aspelund, T. Bhangale, G. Boucher, D. I. Chasman, C. Chen, L. Coin, M. N. Cooper, A. L. Dixon, Q. Gibson, E. Grundberg, K. Hao, M. Juhani Junttila, L. M. Kaplan, J. Kettunen, I. R. Konig, T. Kwan, R. W. Lawrence, D. F. Levinson, M. Lorentzon, B. McKnight, A. P. Morris, M. Muller, J. Suh Ngwa, S. Purcell, S. Rafelt, R. M. Salem, E. Salvi, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, 467(7317):832–8, 2010.

[129] S. Lee, G. R. Abecasis, M. Boehnke, and X. Lin. Rare-variant association analysis: study designs and statistical tests. *Am J Hum Genet*, 95(1):5–23, 2014.

[130] S. Lee, M. J. Emond, M. J. Bamshad, K. C. Barnes, M. J. Rieder, D. A. Nickerson, D. C. Christiani, M. M. Wurfel, and X. Lin. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet*, 91(2):224–37, 2012.

[131] M. Lek, K. J. Karczewski, E. V. Minikel, K. E. Samocha, E. Banks, T. Fennell, A. H. O'Donnell-Luria, J. S. Ware, A. J. Hill, B. B. Cummings, T. Tukiainen,

D. P. Birnbaum, J. A. Kosmicki, L. E. Duncan, K. Estrada, F. Zhao, J. Zou, E. Pierce-Hoffman, J. Berghout, D. N. Cooper, N. Deflaux, M. DePristo, R. Do, J. Flannick, M. Fromer, L. Gauthier, J. Goldstein, N. Gupta, D. Howrigan, A. Kiezun, M. I. Kurki, A. L. Moonshine, P. Natarajan, L. Orozco, G. M. Peloso, R. Poplin, M. A. Rivas, V. Ruano-Rubio, S. A. Rose, D. M. Ruderfer, K. Shakir, P. D. Stenson, C. Stevens, B. P. Thomas, G. Tiao, M. T. Tusie-Luna, B. Weisburd, H. H. Won, D. Yu, D. M. Altshuler, D. Ardissino, M. Boehnke, J. Danesh, S. Donnelly, R. Elosua, J. C. Florez, S. B. Gabriel, G. Getz, S. J. Glatt, C. M. Hultman, S. Kathiresan, M. Laakso, S. McCarroll, M. I. McCarthy, D. McGovern, R. McPherson, B. M. Neale, A. Palotie, S. M. Purcell, D. Saleheen, J. M. Scharf, P. Sklar, P. F. Sullivan, J. Tuomilehto, M. T. Tsuang, H. C. Watkins, J. G. Wilson, M. J. Daly, D. G. MacArthur, and Consortium Exome Aggregation. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616):285–91, 2016.

[132] E. B. Lewis and F. Bacher. Methods of feeding ethyl methane sulfonate (ems) to drosophila males. *Drosophila Inf. Service*, 43(193), 1968.

[133] C. Li, M. Li, J. R. Long, Q. Cai, and W. Zheng. Evaluating cost efficiency of snp chips in genome-wide association studies. *Genet Epidemiol*, 32(5):387–95, 2008.

[134] H. Li. Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. *arXiv*, 2013.

[135] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–9, 2009.

[136] Y. Li and M. Kellis. Joint bayesian inference of risk variants and tissue-specific epigenomic enrichments across multiple complex human diseases. *Nucleic Acids Res*, 44(18):e144, 2016.

[137] P. Liu, N. A. Jenkins, and N. G. Copeland. A highly efficient recombineering-based method for generating conditional knockout mutations. *Genome Res*, 13(3):476–84, 2003.

[138] A. E. Locke, B. Kahali, S. I. Berndt, A. E. Justice, T. H. Pers, F. R. Day, C. Powell, S. Vedantam, M. L. Buchkovich, J. Yang, D. C. Croteau-Chonka, T. Esko, T. Fall, T. Ferreira, S. Gustafsson, Z. Kutalik, J. Luan, R. Magi, J. C. Randall, T. W. Winkler, A. R. Wood, T. Workalemahu, J. D. Faul, J. A.

Smith, J. H. Zhao, W. Zhao, J. Chen, R. Fehrmann, A. K. Hedman, J. Karjalainen, E. M. Schmidt, D. Absher, N. Amin, D. Anderson, M. Beekman, J. L. Bolton, J. L. Bragg-Gresham, S. Buyske, A. Demirkan, G. Deng, G. B. Ehret, B. Feenstra, M. F. Feitosa, K. Fischer, A. Goel, J. Gong, A. U. Jackson, S. Kanoni, M. E. Kleber, K. Kristiansson, U. Lim, V. Lotay, M. Mangino, I. M. Leach, C. Medina-Gomez, S. E. Medland, M. A. Nalls, C. D. Palmer, D. Pasko, S. Pechlivanis, M. J. Peters, I. Prokopenko, D. Shungin, A. Stancakova, R. J. Strawbridge, Y. J. Sung, T. Tanaka, A. Teumer, S. Trompet, S. W. van der Laan, J. van Setten, J. V. Van Vliet-Ostaptchouk, Z. Wang, L. Yengo, W. Zhang, A. Isaacs, E. Albrecht, J. Arnlov, G. M. Arscott, A. P. Attwood, S. Bandinelli, A. Barrett, I. N. Bas, C. Bellis, A. J. Bennett, C. Berne, R. Blagieva, M. Bluher, S. Bohringer, L. L. Bonnycastle, Y. Bottcher, H. A. Boyd, M. Bruinenberg, I. H. Caspersen, Y. I. Chen, R. Clarke, E. W. Daw, A. J. M. de Craen, G. Delgado, M. Dimitriou, et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 518(7538):197–206, 2015.

[139] K. E. Lohmueller, A.R. Indap, S. Schmidt, A.R. Boyko, R.D. Hernandez, M.J. Hubisz, J.J. Sninsky, T.J. White, S.R. Sunyaev, R. Nielsen, et al. Proportionally more deleterious genetic variation in European than in African populations. *Nature*, 451(7181):994–997, 2008.

[140] K. E. Lohmueller, C. L. Pearce, M. Pike, E. S. Lander, and J. N. Hirschhorn. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat Genet*, 33(2):177–82, 2003.

[141] M. Lynch, J. Conery, and R. Burger. Mutational meltdowns in sexual populations. *Evolution*, pages 1067–1080, 1995.

[142] G. Maartens, C. Celum, and S. R. Lewin. Hiv infection: epidemiology, pathogenesis, treatment, and prevention. *Lancet*, 384(9939):258–71, 2014.

[143] J. MacArthur, E. Bowler, M. Cerezo, L. Gil, P. Hall, E. Hastings, H. Junkins, A. McMahon, A. Milano, J. Morales, Z. M. Pendlington, D. Welter, T. Burdett, L. Hindorff, P. Flicek, F. Cunningham, and H. Parkinson. The new nhgri-ebi catalog of published genome-wide association studies (gwas catalog). *Nucleic Acids Res*, 45(D1):D896–D901, 2017.

[144] B. Maher. Personal genomes: The case of the missing heritability. *Nature*, 456(7218):18–21, 2008.

[145] T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. Mackay, S. A. McCarroll, and P. M. Visscher. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–53, 2009.

[146] Teri A Manolio, Francis S Collins, Nancy J Cox, David B Goldstein, Lucia A Hindorff, David J Hunter, Mark I McCarthy, Erin M Ramos, Lon R Cardon, Aravinda Chakravarti, et al. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, 2009.

[147] J. Marchini, B. Howie, S. Myers, G. McVean, and P. Donnelly. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet*, 39(7):906–13, 2007.

[148] G. T. Marth, F. Yu, A.R. Indap, K. Garimella, S. Gravel, W.F. Leong, C. Tyler-Smith, M. Bainbridge, T. Blackwell, X. Zheng-Bradley, et al. The functional spectrum of low-frequency coding variation. *Genome Biology*, 12(9):R84, 2011.

[149] E. R. Martin, J. R. Gilbert, E. H. Lai, J. Riley, A. R. Rogala, B. D. Slotterbeck, C. A. Sipe, J. M. Grubber, L. L. Warren, P. M. Conneally, A. M. Saunders, D. E. Schmechel, I. Purvis, M. A. Pericak-Vance, A. D. Roses, and J. M. Vance. Analysis of association at single nucleotide polymorphisms in the apoe region. *Genomics*, 63(1):7–12, 2000.

[150] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo. The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome Res*, 20(9):1297–303, 2010.

[151] P. J. McLaren, C. Coulonges, S. Ripke, L. van den Berg, S. Buchbinder, M. Carrington, A. Cossarizza, J. Dalmau, S. G. Deeks, O. Delaneau, A. De Luca, J. J. Goedert, D. Haas, J. T. Herbeck, S. Kathiresan, G. D. Kirk, O. Lambotte, M. Luo, S. Mallal, D. van Manen, J. Martinez-Picado, L. Meyer, J. M. Miro, J. I. Mullins, N. Obel, S. J. O'Brien, F. Pereyra, F. A. Plummer, G. Poli, Y. Qi, P. Rucart, M. S. Sandhu, P. R. Shea, H. Schuitemaker, I. Theodorou, F. Vannberg, J. Veldink, B. D. Walker, A. Weintrob, C. A. Winkler, S. Wolinsky, A. Telenti, D. B. Goldstein, P. I. de Bakker, J. F. Zagury, and J. Fellay.

Association study of common genetic variants and hiv-1 acquisition in 6,300 infected cases and 7,200 controls. *PLoS Pathog*, 9(7):e1003515, 2013.

[152] Graham McVicker, David Gordon, Colleen Davis, and Phil Green. Widespread genomic signatures of natural selection in hominid evolution. *PLoS genetics*, 5(5):e1000471, 2009.

[153] Y. Momozawa, M. Mni, K. Nakamura, W. Coppieters, S. Almer, L. Amininejad, I. Cleynen, J. F. Colombel, P. de Rijk, O. Dewit, Y. Finkel, M. A. Gassull, D. Goossens, D. Laukens, M. Lemann, C. Libioulle, C. O'Morain, C. Reenaers, P. Rutgeerts, C. Tysk, D. Zelenika, M. Lathrop, J. Del-Favero, J. P. Hugot, M. de Vos, D. Franchimont, S. Vermeire, E. Louis, and M. Georges. Resequencing of positional candidates identifies low frequency il23r coding variants protecting against inflammatory bowel disease. *Nat Genet*, 43(1):43–7, 2011.

[154] P. Moorjani, Z. Gao, and M. Przeworski. Human germline mutation and the erratic evolutionary clock. *PLoS Biol*, 14(10):e2000744, 2016.

[155] M. Morley, C. M. Molony, T. M. Weber, J. L. Devlin, K. G. Ewens, R. S. Spielman, and V. G. Cheung. Genetic analysis of genome-wide variation in human gene expression. *Nature*, 430(7001):743–7, 2004.

[156] B. M. Neale, Y. Kou, L. Liu, A. Ma'ayan, K. E. Samocha, A. Sabo, C. F. Lin, C. Stevens, L. S. Wang, V. Makarov, P. Polak, S. Yoon, J. Maguire, E. L. Crawford, N. G. Campbell, E. T. Geller, O. Valladares, C. Schafer, H. Liu, T. Zhao, G. Cai, J. Lihm, R. Dannenfelser, O. Jabado, Z. Peralta, U. Nagaswamy, D. Muzny, J. G. Reid, I. Newsham, Y. Wu, L. Lewis, Y. Han, B. F. Voight, E. Lim, E. Rossin, A. Kirby, J. Flannick, M. Fromer, K. Shakir, T. Fennell, K. Garimella, E. Banks, R. Poplin, S. Gabriel, M. DePristo, J. R. Wimbish, B. E. Boone, S. E. Levy, C. Betancur, S. Sunyaev, E. Boerwinkle, J. D. Buxbaum, Jr. Cook, E. H., B. Devlin, R. A. Gibbs, K. Roeder, G. D. Schellenberg, J. S. Sutcliffe, and M. J. Daly. Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature*, 485(7397):242–5, 2012.

[157] M. R. Nelson, D. Wegmann, M. G. Ehm, D. Kessner, P. St Jean, C. Verzilli, J. Shen, Z. Tang, S. A. Bacanu, D. Fraser, L. Warren, J. Aponte, M. Zawistowski, X. Liu, H. Zhang, Y. Zhang, J. Li, Y. Li, L. Li, P. Woollard, S. Topp, M. D. Hall, K. Nangle, J. Wang, G. Abecasis, L. R. Cardon, S. Zollner, J. C. Whittaker, S. L. Chissoe, J. Novembre, and V. Mooser. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science*, 337(6090):100–4, 2012.

[158] M. R. Nelson, D. Wegmann, M.G. Ehm, D. Kessner, P.S. Jean, C. Verzilli, J. Shen, Z. Tang, S.A. Bacanu, D. Fraser, et al. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science*, 337(6090):100–104, 2012.

[159] C. Nowak, S. Salihovic, A. Ganna, S. Brandmaier, T. Tukiainen, C. D. Broeckling, P. K. Magnusson, J. E. Prenni, R. Wang-Sattler, A. Peters, K. Strauch, T. Meitinger, V. Giedraitis, J. Arnlov, C. Berne, C. Gieger, S. Ripatti, L. Lind, N. L. Pedersen, J. Sundstrom, E. Ingelsson, and T. Fall. Effect of insulin resistance on monounsaturated fatty acid levels: A multi-cohort non-targeted metabolomics and mendelian randomization study. *PLoS Genet*, 12(10):e1006379, 2016.

[160] S. J. O'Brien and S. L. Hendrickson. Host genomic influences on hiv/aids. *Genome Biol*, 14(1):201, 2013.

[161] M. C. O'Donovan and M. J. Owen. Candidate-gene association studies of schizophrenia. *Am J Hum Genet*, 65(3):587–92, 1999.

[162] P. F. O'Reilly, C. J. Hoggart, Y. Pomyen, F. C. Calboli, P. Elliott, M. R. Jarvelin, and L. J. Coin. Multiphen: joint model of multiple phenotypes can increase discovery in gwas. *PLoS One*, 7(5):e34861, 2012.

[163] S. P. Otto and M.C. Whitlock. The probability of fixation in populations of changing size. *Genetics*, 146(2):723–733, 1997.

[164] A. A. Pai, J. K. Pritchard, and Y. Gilad. The genetic and mechanistic basis for variation in gene regulation. *PLoS Genet*, 11(1):e1004857, 2015.

[165] B. Pasaniuc and A. L. Price. Dissecting the genetics of complex traits using summary association statistics. *Nat Rev Genet*, 18(2):117–127, 2017.

[166] B. Pasche and N. Yi. Candidate gene association studies: successes and failures. *Curr Opin Genet Dev*, 20(3):257–61, 2010.

[167] I. Pe'er, P. I. de Bakker, J. Maller, R. Yelensky, D. Altshuler, and M. J. Daly. Evaluating and improving power in whole-genome association studies using fixed marker sets. *Nat Genet*, 38(6):663–7, 2006.

[168] K. Pelak, D. B. Goldstein, N. M. Walley, J. Fellay, D. Ge, K. V. Shianna, C. Gumbs, X. Gao, J. M. Maia, K. D. Cronin, S. K. Hussain, M. Carrington,

N. L. Michael, A. C. Weintrob, H. I. V. Working Group Infectious Disease Clinical Research Program, Allergy National Institute of, and H. I. V. Aids Vaccine Immunology Infectious Diseases Center for. Host determinants of hiv-1 control in african americans. *J Infect Dis*, 201(8):1141–9, 2010.

[169] F. Pereyra, X. Jia, P. J. McLaren, A. Telenti, P. I. de Bakker, B. D. Walker, S. Ripke, C. J. Brumme, S. L. Pulit, M. Carrington, C. M. Kadie, J. M. Carlson, D. Heckerman, R. R. Graham, R. M. Plenge, S. G. Deeks, L. Gianniny, G. Crawford, J. Sullivan, E. Gonzalez, L. Davies, A. Camargo, J. M. Moore, N. Beattie, S. Gupta, A. Crenshaw, N. P. Burtt, C. Guiducci, N. Gupta, X. Gao, Y. Qi, Y. Yuki, A. Piechocka-Trocha, E. Cutrell, R. Rosenberg, K. L. Moss, P. Lemay, J. O'Leary, T. Schaefer, P. Verma, I. Toth, B. Block, B. Baker, A. Rothchild, J. Lian, J. Proudfoot, D. M. Alvino, S. Vine, M. M. Addo, T. M. Allen, M. Altfeld, M. R. Henn, S. Le Gall, H. Streeck, D. W. Haas, D. R. Kuritzkes, G. K. Robbins, R. W. Shafer, R. M. Gulick, C. M. Shikuma, R. Haubrich, S. Riddler, P. E. Sax, E. S. Daar, H. J. Ribaudo, B. Agan, S. Agarwal, R. L. Ahern, B. L. Allen, S. Altidor, E. L. Altschuler, S. Ambardar, K. Anastos, B. Anderson, V. Anderson, U. Andrady, D. Antoniskis, D. Bangsberg, D. Barbaro, W. Barrie, J. Bartczak, S. Barton, P. Basden, N. Basgoz, S. Bazner, N. C. Bellos, A. M. Benson, J. Berger, N. F. Bernard, A. M. Bernard, C. Birch, S. J. Bodner, R. K. Bolan, E. T. Boudreaux, M. Bradley, J. F. Braun, J. E. Brndjar, S. J. Brown, K. Brown, S. T. Brown, et al. The major genetic determinants of hiv-1 control affect hla class i peptide presentation. *Science*, 330(6010):1551–7, 2010.

[170] B. S. Petersen, B. Fredrich, M. P. Hoeppner, D. Ellinghaus, and A. Franke. Opportunities and challenges of whole-genome and -exome sequencing. *BMC Genet*, 18(1):14, 2017.

[171] J. K. Pickrell, T. Berisa, J. Z. Liu, L. Segurel, J. Y. Tung, and D. A. Hinds. Detection and interpretation of shared genetic influences on 42 human traits. *Nat Genet*, 48(7):709–17, 2016.

[172] H. F. Porter and P. F. O'Reilly. Multivariate simulation framework reveals performance of multi-trait gwas methods. *Sci Rep*, 7:38837, 2017.

[173] A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*, 38(8):904–9, 2006.

[174] A. L. Price, C. C. Spencer, and P. Donnelly. Progress and promise in understanding the genetic basis of common diseases. *Proc Biol Sci*, 282(1821):20151684, 2015.

[175] J. K. Pritchard. Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet*, 69(1):124–37, 2001.

[176] J. K. Pritchard. Are rare variants responsible for susceptibility to complex diseases? *The American Journal of Human Genetics*, 69(1):124–137, 2001.

[177] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly, and P. C. Sham. Plink: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, 81(3):559–75, 2007.

[178] S. M. Purcell, J. L. Moran, M. Fromer, D. Ruderfer, N. Solovieff, P. Roussos, C. O'Dushlaine, K. Chambert, S. E. Bergen, A. Kahler, L. Duncan, E. Stahl, G. Genovese, E. Fernandez, M. O. Collins, N. H. Komiyama, J. S. Choudhary, P. K. Magnusson, E. Banks, K. Shakir, K. Garimella, T. Fennell, M. DePristo, S. G. Grant, S. J. Haggarty, S. Gabriel, E. M. Scolnick, E. S. Lander, C. M. Hultman, P. F. Sullivan, S. A. McCarroll, and P. Sklar. A polygenic burden of rare disruptive mutations in schizophrenia. *Nature*, 506(7487):185–90, 2014.

[179] S. D. Rees, M. Z. Hydrie, J. P. O'Hare, S. Kumar, A. S. Shera, A. Basit, A. H. Barnett, and M. A. Kelly. Effects of 16 genetic variants on fasting glucose and type 2 diabetes in south asians: Adcy5 and glis3 variants may predispose to type 2 diabetes. *PLoS One*, 6(9):e24710, 2011.

[180] D. E. Reich and E. S. Lander. On the allelic spectrum of human disease. *Trends Genet*, 17(9):502–10, 2001.

[181] N. Risch. Linkage strategies for genetically complex traits. i. multilocus models. *Am J Hum Genet*, 46(2):222–8, 1990.

[182] N. Risch and K. Merikangas. The future of genetic studies of complex human diseases. *Science*, 273(5281):1516–7, 1996.

[183] M. A. Rivas, M. Beaudoin, A. Gardet, C. Stevens, Y. Sharma, C. K. Zhang, G. Boucher, S. Ripke, D. Ellinghaus, N. Burtt, T. Fennell, A. Kirby, A. Latiano, P. Goyette, T. Green, J. Halfvarson, T. Haritunians, J. M. Korn, F. Kuruvilla, C. Lagace, B. Neale, K. S. Lo, P. Schumm, L. Torkvist, Diabetes National

Institute of, Consortium Digestive Kidney Diseases Inflammatory Bowel Disease Genetics, Consortium United Kingdom Inflammatory Bowel Disease Genetics, Consortium International Inflammatory Bowel Disease Genetics, M. C. Dubinsky, S. R. Brant, M. S. Silverberg, R. H. Duerr, D. Altshuler, S. Gabriel, G. Lettre, A. Franke, M. D'Amato, D. P. McGovern, J. H. Cho, J. D. Rioux, R. J. Xavier, and M. J. Daly. Deep resequencing of gwas loci identifies independent rare variants associated with inflammatory bowel disease. *Nat Genet*, 43(11):1066–73, 2011.

[184] Consortium Roadmap Epigenomics, A. Kundaje, W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi, P. Kheradpour, Z. Zhang, J. Wang, M. J. Ziller, V. Amin, J. W. Whitaker, M. D. Schultz, L. D. Ward, A. Sarkar, G. Quon, R. S. Sandstrom, M. L. Eaton, Y. C. Wu, A. R. Pfenning, X. Wang, M. Claussnitzer, Y. Liu, C. Coarfa, R. A. Harris, N. Shoresh, C. B. Epstein, E. Gjoneska, D. Leung, W. Xie, R. D. Hawkins, R. Lister, C. Hong, P. Gascard, A. J. Mungall, R. Moore, E. Chuah, A. Tam, T. K. Canfield, R. S. Hansen, R. Kaul, P. J. Sabo, M. S. Bansal, A. Carles, J. R. Dixon, K. H. Farh, S. Feizi, R. Karlic, A. R. Kim, A. Kulkarni, D. Li, R. Lowdon, G. Elliott, T. R. Mercer, S. J. Neph, V. Onuchic, P. Polak, N. Rajagopal, P. Ray, R. C. Sallari, K. T. Siebenthall, N. A. Sinnott-Armstrong, M. Stevens, R. E. Thurman, J. Wu, B. Zhang, X. Zhou, A. E. Beaudet, L. A. Boyer, P. L. De Jager, P. J. Farnham, S. J. Fisher, D. Haussler, S. J. Jones, W. Li, M. A. Marra, M. T. McManus, S. Sunyaev, J. A. Thomson, T. D. Tlsty, L. H. Tsai, W. Wang, R. A. Waterland, M. Q. Zhang, L. H. Chadwick, B. E. Bernstein, J. F. Costello, J. R. Ecker, M. Hirst, A. Meissner, A. Milosavljevic, B. Ren, J. A. Stamatoyannopoulos, T. Wang, and M. Kellis. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–30, 2015.

[185] J. Roote and A. Prokop. How to design a genetic mating scheme: a basic training package for drosophila genetics. *G3 (Bethesda)*, 3(2):353–8, 2013.

[186] W. L. Russell, E. M. Kelly, P. R. Hunsicker, J. W. Bangham, S. C. Maddux, and E. L. Phipps. Specific-locus test shows ethylnitrosourea to be the most potent mutagen in the mouse. *Proc Natl Acad Sci U S A*, 76(11):5818–9, 1979.

[187] M. Samson, O. Labbe, C. Mollereau, G. Vassart, and M. Parmentier. Molecular cloning and functional expression of a new human cc-chemokine receptor gene. *Biochemistry*, 35(11):3362–7, 1996.

[188] R. Saxena, M. F. Hivert, C. Langenberg, T. Tanaka, J. S. Pankow, P. Vollenweider, V. Lyssenko, N. Bouatia-Naji, J. Dupuis, A. U. Jackson, W. H. Kao, M. Li, N. L. Glazer, A. K. Manning, J. Luan, H. M. Stringham, I. Prokopenko, T. Johnson, N. Grarup, T. W. Boesgaard, C. Lecoeur, P. Shrader, J. O'Connell, E. Ingelsson, D. J. Couper, K. Rice, K. Song, C. H. Andreasen, C. Dina, A. Kottgen, O. Le Bacquer, F. Pattou, J. Taneera, V. Steinthorsdottir, D. Rybin, K. Ardlie, M. Sampson, L. Qi, M. van Hoek, M. N. Weedon, Y. S. Aulchenko, B. F. Voight, H. Grallert, B. Balkau, R. N. Bergman, S. J. Bielinski, A. Bonnefond, L. L. Bonnycastle, K. Borch-Johnsen, Y. Bottcher, E. Brunner, T. A. Buchanan, S. J. Bumpstead, C. Cavalcanti-Proenca, G. Charpentier, Y. D. Chen, P. S. Chines, F. S. Collins, M. Cornelis, J. Crawford G, J. Delplanque, A. Doney, J. M. Egan, M. R. Erdos, M. Firmann, N. G. Forouhi, C. S. Fox, M. O. Goodarzi, J. Graessler, A. Hingorani, B. Isomaa, T. Jorgensen, M. Kivimaki, P. Kovacs, K. Krohn, M. Kumari, T. Lauritzen, C. Levy-Marchal, V. Mayor, J. B. McAteer, D. Meyre, B. D. Mitchell, K. L. Mohlke, M. A. Morken, N. Narisu, C. N. Palmer, R. Pakyz, L. Pascoe, F. Payne, D. Pearson, W. Rathmann, A. Sandbaek, A. A. Sayer, L. J. Scott, S. J. Sharp, E. Sijbrands, A. Singleton, D. S. Siscovick, N. L. Smith, T. Sparso, et al. Genetic variation in gipr influences the glucose and insulin responses to an oral glucose challenge. *Nat Genet*, 42(2):142–8, 2010.

[189] Stephen F Schaffner, Catherine Foo, Stacey Gabriel, David Reich, Mark J Daly, and David Altshuler. Calibrating a coalescent simulation of human genome sequence variation. *Genome Research*, 15(11):1576–1583, 2005.

[190] N. J. Schork, S. S. Murray, K. A. Frazer, and E. J. Topol. Common vs. rare allele hypotheses for complex diseases. *Curr Opin Genet Dev*, 19(3):212–9, 2009.

[191] Jana Marie Schwarz, Christian Rödelsperger, Markus Schuelke, and Dominik Seelow. MutationTaster evaluates disease-causing potential of sequence alterations. *Nature Methods*, 7(8):575–576, 2010.

[192] P. R. Shea, K. V. Shianna, M. Carrington, and D. B. Goldstein. Host genetics of hiv acquisition and viral control. *Annu Rev Med*, 64:203–17, 2013.

[193] J. Shendure and H. Ji. Next-generation dna sequencing. *Nat Biotechnol*, 26(10):1135–45, 2008.

[194] S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin. dbsnp: the ncbi database of genetic variation. *Nucleic Acids Res*, 29(1):308–11, 2001.

[195] H. Shi, G. Kichaev, and B. Pasaniuc. Contrasting the genetic architecture of 30 complex traits from summary association data. *Am J Hum Genet*, 99(1):139–53, 2016.

[196] D. Shriner. Moving toward system genetics through multiple trait analysis in genome-wide association studies. *Front Genet*, 3:1, 2012.

[197] H. A. Shuman and T. J. Silhavy. The art and design of genetic screens: Escherichia coli. *Nat Rev Genet*, 4(6):419–31, 2003.

[198] D. Shungin, T. W. Winkler, D. C. Croteau-Chonka, T. Ferreira, A. E. Locke, R. Magi, R. J. Strawbridge, T. H. Pers, K. Fischer, A. E. Justice, T. Workalemahu, J. M. W. Wu, M. L. Buchkovich, N. L. Heard-Costa, T. S. Roman, A. W. Drong, C. Song, S. Gustafsson, F. R. Day, T. Esko, T. Fall, Z. Kutalik, J. Luan, J. C. Randall, A. Scherag, S. Vedantam, A. R. Wood, J. Chen, R. Fehrmann, J. Karjalainen, B. Kahali, C. T. Liu, E. M. Schmidt, D. Absher, N. Amin, D. Anderson, M. Beekman, J. L. Bragg-Gresham, S. Buyske, A. Demirkan, G. B. Ehret, M. F. Feitosa, A. Goel, A. U. Jackson, T. Johnson, M. E. Kleber, K. Kristiansson, M. Mangino, I. M. Leach, C. Medina-Gomez, C. D. Palmer, D. Pasko, S. Pechlivanis, M. J. Peters, I. Prokopenko, A. Stancakova, Y. J. Sung, T. Tanaka, A. Teumer, J. V. Van Vliet-Ostaptchouk, L. Yengo, W. Zhang, E. Albrecht, J. Arnlov, G. M. Arscott, S. Bandinelli, A. Barrett, C. Bellis, A. J. Bennett, C. Berne, M. Bluher, S. Bohringer, F. Bonnet, Y. Bottcher, M. Bruinenberg, D. B. Carba, I. H. Caspersen, R. Clarke, E. W. Daw, J. Deelen, E. Deelman, G. Delgado, A. S. Doney, N. Eklund, M. R. Erdos, K. Estrada, E. Eury, N. Friedrich, M. E. Garcia, V. Giedraitis, B. Gigante, A. S. Go, A. Golay, H. Grallert, T. B. Grammer, J. Grassler, J. Grewal, C. J. Groves, T. Haller, G. Hallmans, et al. New genetic loci link adipose and insulin biology to body fat distribution. *Nature*, 518(7538):187–196, 2015.

[199] V. Simon, D. D. Ho, and Q. Abdool Karim. Hiv/aids epidemiology, pathogenesis, prevention, and treatment. *Lancet*, 368(9534):489–504, 2006.

[200] M. Slatkin. Epigenetic inheritance and the missing heritability problem. *Genetics*, 182(3):845–50, 2009.

[201] G. D. Smith and S. Ebrahim. 'mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol*, 32(1):1–22, 2003.

[202] N. Solovieff, C. Cotsapas, P. H. Lee, S. M. Purcell, and J. W. Smoller. Pleiotropy in complex traits: challenges and strategies. *Nat Rev Genet*, 14(7):483–95, 2013.

[203] E. K. Speliotes, C. J. Willer, S. I. Berndt, K. L. Monda, G. Thorleifsson, A. U. Jackson, H. L. Allen, C. M. Lindgren, J. Luan, R. Magi, J. C. Randall, S. Vedantam, T. W. Winkler, L. Qi, T. Workalemahu, I. M. Heid, V. Steinthorsdottir, H. M. Stringham, M. N. Weedon, E. Wheeler, A. R. Wood, T. Ferreira, R. J. Weyant, A. V. Segre, K. Estrada, L. Liang, J. Nemesh, J. H. Park, S. Gustafsson, T. O. Kilpelainen, J. Yang, N. Bouatia-Naji, T. Esko, M. F. Feitosa, Z. Kutalik, M. Mangino, S. Raychaudhuri, A. Scherag, A. V. Smith, R. Welch, J. H. Zhao, K. K. Aben, D. M. Absher, N. Amin, A. L. Dixon, E. Fisher, N. L. Glazer, M. E. Goddard, N. L. Heard-Costa, V. Hoesel, J. J. Hottenga, A. Johansson, T. Johnson, S. Ketkar, C. Lamina, S. Li, M. F. Moffatt, R. H. Myers, N. Narisu, J. R. Perry, M. J. Peters, M. Preuss, S. Ripatti, F. Rivadeneira, C. Sandholt, L. J. Scott, N. J. Timpson, J. P. Tyrer, S. van Wingerden, R. M. Watanabe, C. C. White, F. Wiklund, C. Barlassina, D. I. Chasman, M. N. Cooper, J. O. Jansson, R. W. Lawrence, N. Pellikka, I. Prokopenko, J. Shi, E. Thiering, H. Alavere, M. T. Alibrandi, P. Almgren, A. M. Arnold, T. Aspelund, L. D. Atwood, B. Balkau, A. J. Balmforth, A. J. Bennett, Y. Ben-Shlomo, R. N. Bergman, S. Bergmann, H. Biebermann, A. I. Blakemore, T. Boes, L. L. Bonnycastle, S. R. Bornstein, M. J. Brown, T. A. Buchanan, et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat Genet*, 42(11):937–48, 2010.

[204] C. C. Spencer, Z. Su, P. Donnelly, and J. Marchini. Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS Genet*, 5(5):e1000477, 2009.

[205] J. Stappert, J. Wirsching, and R. Kemler. A pcr method for introducing mutations into cloned dna by joining an internal primer to a tagged flanking primer. *Nucleic Acids Res*, 20(3):624, 1992.

[206] G. R. Stark and A. V. Gudkov. Forward genetics in mammalian cells: functional approaches to gene discovery. *Hum Mol Genet*, 8(10):1925–38, 1999.

[207] M. Stephens. A unified framework for association analysis with multiple related phenotypes. *PLoS One*, 8(7):e65245, 2013.

[208] M. Stephens. False discovery rates: a new deal. *Biostatistics*, 18(2):275–294, 2016.

[209] M. Stremlau, C. M. Owens, M. J. Perron, M. Kiessling, P. Autissier, and J. Sodroski. The cytoplasmic body component trim5alpha restricts hiv-1 infection in old world monkeys. *Nature*, 427(6977):848–53, 2004.

[210] S. H. Stricker, A. Koferle, and S. Beck. From profiles to function in epigenomics. *Nat Rev Genet*, 18(1):51–66, 2017.

[211] C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray, B. Liu, P. Matthews, G. Ong, J. Pell, A. Silman, A. Young, T. Sprosen, T. Peakman, and R. Collins. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med*, 12(3):e1001779, 2015.

[212] J. X. Sun, A. Helgason, G. Masson, S. S. Ebenesersdottir, H. Li, S. Mallick, S. Gnerre, N. Patterson, A. Kong, D. Reich, and K. Stefansson. A direct characterization of human mutation based on microsatellites. *Nat Genet*, 44(10):1161–5, 2012.

[213] A. Suzuki, R. Yamada, X. Chang, S. Tokuhiro, T. Sawada, M. Suzuki, M. Nagasaki, M. Nakayama-Hamada, R. Kawaida, M. Ono, M. Ohtsuki, H. Furukawa, S. Yoshino, M. Yukioka, S. Tohma, T. Matsubara, S. Wakitani, R. Teshima, Y. Nishioka, A. Sekine, A. Iida, A. Takahashi, T. Tsunoda, Y. Nakamura, and K. Yamamoto. Functional haplotypes of padi4, encoding citrullinating enzyme peptidylarginine deiminase 4, are associated with rheumatoid arthritis. *Nat Genet*, 34(4):395–402, 2003.

[214] F. Tajima. The effect of change in population size on dna polymorphism. *Genetics*, 123(3):597–601, 1989.

[215] J. A. Tennessen, A. W. Bigham, T.D. O'Connor, W. Fu, E.E. Kenny, S. Gravel, S. McGee, R. Do, X. Liu, G. Jun, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*, 337(6090):64–69, 2012.

[216] T. M. Teslovich, K. Musunuru, A. V. Smith, A. C. Edmondson, I. M. Stylianou, M. Koseki, J. P. Pirruccello, S. Ripatti, D. I. Chasman, C. J. Willer, C. T. Johansen, S. W. Fouchier, A. Isaacs, G. M. Peloso, M. Barbalic, S. L. Ricketts, J. C. Bis, Y. S. Aulchenko, G. Thorleifsson, M. F. Feitosa, J. Chambers, M. Orho-Melander, O. Melander, T. Johnson, X. Li, X. Guo, M. Li, Y. Shin Cho, M. Jin Go, Y. Jin Kim, J. Y. Lee, T. Park, K. Kim, X. Sim, R. Twee-Hee Ong, D. C. Croteau-Chonka, L. A. Lange, J. D. Smith, K. Song, J. Hua Zhao, X. Yuan, J. Luan, C. Lamina, A. Ziegler, W. Zhang, R. Y. Zee, A. F. Wright, J. C. Witteman, J. F. Wilson, G. Willemsen, H. E. Wichmann, J. B. Whitfield, D. M. Waterworth, N. J. Wareham, G. Waeber, P. Vollenweider, B. F. Voight, V. Vitart, A. G. Uitterlinden, M. Uda, J. Tuomilehto, J. R. Thompson, T. Tanaka, I. Surakka, H. M. Stringham, T. D. Spector, N. Soranzo, J. H. Smit, J. Sinisalo, K. Silander, E. J. Sijbrands, A. Scuteri, J. Scott, D. Schlessinger, S. Sanna, V. Salomaa, J. Saharinen, C. Sabatti, A. Ruokonen, I. Rudan, L. M. Rose, R. Roberts, M. Rieder, B. M. Psaty, P. P. Pramstaller, I. Pichler, M. Perola, B. W. Penninx, N. L. Pedersen, C. Pattaro, A. N. Parker, G. Pare, B. A. Oostra, C. J. O'Donnell, M. S. Nieminen, D. A. Nickerson, G. W. Montgomery, T. Meitinger, R. McPherson, M. I. McCarthy, et al. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, 466(7307):707–13, 2010.

[217] The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, 2010.

[218] Kevin R Thornton, Andrew J Foran, and Anthony D Long. Properties and modeling of GWAS when complex disease risk is due to non-complementing, deleterious mutations in genes of large effect. *PLoS Genetics*, 9(2):e1003258, 2013.

[219] MJ Travis, Tamara Münkemüller, Olivia J Burton, Alex Best, Calvin Dytham, and Karin Johst. Deleterious mutations can surf to high densities on the wave front of an expanding population. *Molecular Biology and Evolution*, 24(10):2334–2343, 2007.

[220] S. M. Urbut, G. Wang, and M. Stephens. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *bioRxiv*, 2016.

[221] W. Valdar, L. C. Solberg, D. Gauguier, S. Burnett, P. Klenerman, W. O. Cookson, M. S. Taylor, J. N. Rawlins, R. Mott, and J. Flint. Genome-wide

genetic association of complex traits in heterogeneous stock mice. *Nat Genet*, 38(8):879–87, 2006.

[222] E. J. Vallender and B. T. Lahn. Positive selection on the human genome. *Hum Mol Genet*, 13 Spec No 2:R245–54, 2004.

[223] N. Van Damme, D. Goff, C. Katsura, R. L. Jorgenson, R. Mitchell, M. C. Johnson, E. B. Stephens, and J. Guatelli. The interferon-induced protein bst-2 restricts hiv-1 release and is downregulated from the cell surface by the viral vpu protein. *Cell Host Microbe*, 3(4):245–52, 2008.

[224] G. A. Van der Auwera, M. O. Carneiro, C. Hartl, R. Poplin, G. Del Angel, A. Levy-Moonshine, T. Jordan, K. Shakir, D. Roazen, J. Thibault, E. Banks, K. V. Garimella, D. Altshuler, S. Gabriel, and M. A. DePristo. From fastq data to high confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr Protoc Bioinformatics*, 43:11 10 1–33, 2013.

[225] P. van der Harst, W. Zhang, I. Mateo Leach, A. Rendon, N. Verweij, J. Sehmi, D. S. Paul, U. Elling, H. Allayee, X. Li, A. Radhakrishnan, S. T. Tan, K. Voss, C. X. Weichenberger, C. A. Albers, A. Al-Hussani, F. W. Asselbergs, M. Ciullo, F. Danjou, C. Dina, T. Esko, D. M. Evans, L. Franke, M. Gogele, J. Hartiala, M. Hersch, H. Holm, J. J. Hottenga, S. Kanoni, M. E. Kleber, V. Lagou, C. Langenberg, L. M. Lopez, L. P. Lyytikainen, O. Melander, F. Murgia, I. M. Nolte, P. F. O'Reilly, S. Padmanabhan, A. Parsa, N. Pirastu, E. Porcu, L. Portas, I. Prokopenko, J. S. Ried, S. Y. Shin, C. S. Tang, A. Teumer, M. Traglia, S. Ulivi, H. J. Westra, J. Yang, J. H. Zhao, F. Anni, A. Abdellaoui, A. Attwood, B. Balkau, S. Bandinelli, F. Bastardot, B. Benyamin, B. O. Boehm, W. O. Cookson, D. Das, P. I. de Bakker, R. A. de Boer, E. J. de Geus, M. H. de Moor, M. Dimitriou, F. S. Domingues, A. Doring, G. Engstrom, G. I. Eyjolfsson, L. Ferrucci, K. Fischer, R. Galanello, S. F. Garner, B. Genser, Q. D. Gibson, G. Girotto, D. F. Gudbjartsson, S. E. Harris, A. L. Hartikainen, C. E. Hastie, B. Hedblad, T. Illig, J. Jolley, M. Kahonen, I. P. Kema, J. P. Kemp, L. Liang, H. Lloyd-Jones, R. J. Loos, S. Meacham, S. E. Medland, C. Meisinger, Y. Memari, E. Mihailov, K. Miller, M. F. Moffatt, M. Nauck, et al. Seventy-five genetic loci influencing the human red blood cell. *Nature*, 492(7429):369–75, 2012.

[226] P. M. Visscher, M. A. Brown, M. I. McCarthy, and J. Yang. Five years of gwas discovery. *Am J Hum Genet*, 90(1):7–24, 2012.

[227] P. M. Visscher, N. R. Wray, Q. Zhang, P. Sklar, M. I. McCarthy, M. A. Brown, and J. Yang. 10 years of gwas discovery: Biology, function, and translation. *Am J Hum Genet*, 101(1):5–22, 2017.

[228] B. F. Voight, G. M. Peloso, M. Orho-Melander, R. Frikke-Schmidt, M. Barbalic, M. K. Jensen, G. Hindy, H. Holm, E. L. Ding, T. Johnson, H. Schunkert, N. J. Samani, R. Clarke, J. C. Hopewell, J. F. Thompson, M. Li, G. Thorleifsson, C. Newton-Cheh, K. Musunuru, J. P. Pirruccello, D. Saleheen, L. Chen, A. Stewart, A. Schillert, U. Thorsteinsdottir, G. Thorgeirsson, S. Anand, J. C. Engert, T. Morgan, J. Spertus, M. Stoll, K. Berger, N. Martinelli, D. Girelli, P. P. McKeown, C. C. Patterson, S. E. Epstein, J. Devaney, M. S. Burnett, V. Mooser, S. Ripatti, I. Surakka, M. S. Nieminen, J. Sinisalo, M. L. Lokki, M. Perola, A. Havulinna, U. de Faire, B. Gigante, E. Ingelsson, T. Zeller, P. Wild, P. I. de Bakker, O. H. Klungel, A. H. Maitland-van der Zee, B. J. Peters, A. de Boer, D. E. Grobbee, P. W. Kamphuisen, V. H. Deneer, C. C. Elbers, N. C. Onland-Moret, M. H. Hofker, C. Wijmenga, W. M. Verschuren, J. M. Boer, Y. T. van der Schouw, A. Rasheed, P. Frossard, S. Demissie, C. Willer, R. Do, J. M. Ordovas, G. R. Abecasis, M. Boehnke, K. L. Mohlke, M. J. Daly, C. Guiducci, N. P. Burtt, A. Surti, E. Gonzalez, S. Purcell, S. Gabriel, J. Marrugat, J. Peden, J. Erdmann, P. Diemert, C. Willenborg, I. R. Konig, M. Fischer, C. Hengstenberg, A. Ziegler, I. Buysschaert, D. Lambrechts, F. Van de Werf, K. A. Fox, N. E. El Mokhtari, D. Rubin, J. Schrezenmeir, S. Schreiber, et al. Plasma hdl cholesterol and risk of myocardial infarction: a mendelian randomisation study. *Lancet*, 380(9841):572–80, 2012.

[229] Benjamin F Voight, Alison M Adams, Linda A Frisse, Yudong Qian, Richard R Hudson, and Anna Di Rienzo. Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. *Proceedings of the National Academy of Sciences of the United States of America*, 102(51):18508–18513, 2005.

[230] L. V. Wain, G. C. Verwoert, P. F. O'Reilly, G. Shi, T. Johnson, A. D. Johnson, M. Bochud, K. M. Rice, P. Henneman, A. V. Smith, G. B. Ehret, N. Amin, M. G. Larson, V. Mooser, D. Hadley, M. Dorr, J. C. Bis, T. Aspelund, T. Esko, A. C. Janssens, J. H. Zhao, S. Heath, M. Laan, J. Fu, G. Pistis, J. Luan, P. Arora, G. Lucas, N. Pirastu, I. Pichler, A. U. Jackson, R. J. Webster, F. Zhang, J. F. Peden, H. Schmidt, T. Tanaka, H. Campbell, W. Igl, Y. Milaneschi, J. J. Hottenga, V. Vitart, D. I. Chasman, S. Trompet, J. L. Bragg-Gresham, B. Z. Alizadeh, J. C. Chambers, X. Guo, T. Lehtimaki, B. Kuhnel,

L. M. Lopez, O. Polasek, M. Boban, C. P. Nelson, A. C. Morrison, V. Pihur, S. K. Ganesh, A. Hofman, S. Kundu, F. U. Mattace-Raso, F. Rivadeneira, E. J. Sijbrands, A. G. Uitterlinden, S. J. Hwang, R. S. Vasan, T. J. Wang, S. Bergmann, P. Vollenweider, G. Waeber, J. Laitinen, A. Pouta, P. Zitting, W. L. McArdle, H. K. Kroemer, U. Volker, H. Volzke, N. L. Glazer, K. D. Taylor, T. B. Harris, H. Alavere, T. Haller, A. Keis, M. L. Tammesoo, Y. Aulchenko, I. Barroso, K. T. Khaw, P. Galan, S. Hercberg, M. Lathrop, S. Eyheramendy, E. Org, S. Sober, X. Lu, I. M. Nolte, B. W. Penninx, T. Corre, C. Masciullo, C. Sala, L. Groop, B. F. Voight, O. Melander, et al. Genome-wide association study identifies six new loci influencing pulse pressure and mean arterial pressure. *Nat Genet*, 43(10):1005–11, 2011.

[231] J. Wakeley. Coalescent theory: An introduction. 2008.

[232] J. D. Wall and J. K. Pritchard. Haplotype blocks and linkage disequilibrium in the human genome. *Nat Rev Genet*, 4(8):587–97, 2003.

[233] J.D. Wall and M. Przeworski. When did the human population size start increasing? *Genetics*, 155(4):1865–1874, 2000.

[234] K. Wang, M. Li, and H. Hakonarson. Annovar: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*, 38(16):e164, 2010.

[235] Kai Wang, Mingyao Li, and Hakon Hakonarson. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research*, 38(16):e164–e164, 2010.

[236] Q. Wang, Q. Lu, and H. Zhao. A review of study designs and statistical methods for genomic epidemiology studies using next generation sequencing. *Front Genet*, 6:149, 2015.

[237] Consortium Wellcome Trust Case Control. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–78, 2007.

[238] X. Wen, R. Pique-Regi, and F. Luca. Integrating molecular qtl data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization. *PLoS Genet*, 13(3):e1006646, 2017.

[239] C. J. Willer and K. L. Mohlke. Finding genes and variants for lipid levels after genome-wide association analysis. *Curr Opin Lipidol*, 23(2):98–103, 2012.

[240] C. J. Willer, E. M. Schmidt, S. Sengupta, G. M. Peloso, S. Gustafsson, S. Kanoni, A. Ganna, J. Chen, M. L. Buchkovich, S. Mora, J. S. Beckmann, J. L. Bragg-Gresham, H. Y. Chang, A. Demirkan, H. M. Den Hertog, R. Do, L. A. Donnelly, G. B. Ehret, T. Esko, M. F. Feitosa, T. Ferreira, K. Fischer, P. Fontanillas, R. M. Fraser, D. F. Freitag, D. Gurdasani, K. Heikkila, E. Hypponen, A. Isaacs, A. U. Jackson, A. Johansson, T. Johnson, M. Kaakinen, J. Kettunen, M. E. Kleber, X. Li, J. Luan, L. P. Lyytikainen, P. K. E. Magnusson, M. Mangino, E. Mihailov, M. E. Montasser, M. Muller-Nurasyid, I. M. Nolte, J. R. O'Connell, C. D. Palmer, M. Perola, A. K. Petersen, S. Sanna, R. Saxena, S. K. Service, S. Shah, D. Shungin, C. Sidore, C. Song, R. J. Strawbridge, I. Surakka, T. Tanaka, T. M. Teslovich, G. Thorleifsson, E. G. Van den Herik, B. F. Voight, K. A. Volcik, L. L. Waite, A. Wong, Y. Wu, W. Zhang, D. Absher, G. Asiki, I. Barroso, L. F. Been, J. L. Bolton, L. L. Bonnycastle, P. Brambilla, M. S. Burnett, G. Cesana, M. Dimitriou, A. S. F. Doney, A. Doring, P. Elliott, S. E. Epstein, G. Ingi Eyjolfsson, B. Gigante, M. O. Goodarzi, H. Grallert, M. L. Gravito, C. J. Groves, G. Hallmans, A. L. Hartikainen, C. Hayward, D. Hernandez, A. A. Hicks, H. Holm, Y. J. Hung, T. Illig, M. R. Jones, P. Kaleebu, J. J. P. Kastelein, K. T. Khaw, E. Kim, et al. Discovery and refinement of loci associated with lipid levels. *Nat Genet*, 45(11):1274–1283, 2013.

[241] A. R. Wood, T. Esko, J. Yang, S. Vedantam, T. H. Pers, S. Gustafsson, A. Y. Chu, K. Estrada, J. Luan, Z. Kutalik, N. Amin, M. L. Buchkovich, D. C. Croteau-Chonka, F. R. Day, Y. Duan, T. Fall, R. Fehrmann, T. Ferreira, A. U. Jackson, J. Karjalainen, K. S. Lo, A. E. Locke, R. Magi, E. Mihailov, E. Porcu, J. C. Randall, A. Scherag, A. A. Vinkhuyzen, H. J. Westra, T. W. Winkler, T. Workalemahu, J. H. Zhao, D. Absher, E. Albrecht, D. Anderson, J. Baron, M. Beekman, A. Demirkan, G. B. Ehret, B. Feenstra, M. F. Feitosa, K. Fischer, R. M. Fraser, A. Goel, J. Gong, A. E. Justice, S. Kanoni, M. E. Kleber, K. Kristiansson, U. Lim, V. Lotay, J. C. Lui, M. Mangino, I. Mateo Leach, C. Medina-Gomez, M. A. Nalls, D. R. Nyholt, C. D. Palmer, D. Pasko, S. Pechlivanis, I. Prokopenko, J. S. Ried, S. Ripke, D. Shungin, A. Stancakova, R. J. Strawbridge, Y. J. Sung, T. Tanaka, A. Teumer, S. Trompet, S. W. van der Laan, J. van Setten, J. V. Van Vliet-Ostaptchouk, Z. Wang, L. Yengo, W. Zhang, U. Afzal, J. Arnlov, G. M. Arscott, S. Bandinelli, A. Barrett, C. Bellis, A. J. Bennett, C. Berne, M. Bluher, J. L. Bolton, Y. Bottcher, H. A. Boyd, M. Bruinenberg, B. M. Buckley, S. Buyske, I. H. Caspersen, P. S. Chines, R. Clarke, S. Claudi-Boehm, M. Cooper, E. W. Daw, P. A. De Jong, J. Deelen, G. Del-

gado, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet*, 46(11):1173–86, 2014.

[242] M. C. Wu, S. Lee, T. Cai, Y. Li, M. Boehnke, and X. Lin. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet*, 89(1):82–93, 2011.

[243] J. Yang, B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, M. E. Goddard, and P. M. Visscher. Common snps explain a large proportion of the heritability for human height. *Nat Genet*, 42(7):565–9, 2010.

[244] Q. Yang and Y. Wang. Methods for analyzing multivariate phenotypes in genetic association studies. *J Probab Stat*, 2012:652569, 2012.

[245] X. Yuan, D. Waterworth, J. R. Perry, N. Lim, K. Song, J. C. Chambers, W. Zhang, P. Vollenweider, H. Stirnadel, T. Johnson, S. Bergmann, N. D. Beckmann, Y. Li, L. Ferrucci, D. Melzer, D. Hernandez, A. Singleton, J. Scott, P. Elliott, G. Waeber, L. Cardon, T. M. Frayling, J. S. Kooner, and V. Mooser. Population-based genome-wide association studies reveal six loci influencing plasma levels of liver enzymes. *Am J Hum Genet*, 83(4):520–8, 2008.

[246] H. F. Zheng, V. Forgetta, Y. H. Hsu, K. Estrada, A. Rosello-Diez, P. J. Leo, C. L. Dahia, K. H. Park-Min, J. H. Tobias, C. Kooperberg, A. Kleinman, U. Styrkarsdottir, C. T. Liu, C. Uggla, D. S. Evans, C. M. Nielson, K. Walter, U. Pettersson-Kymmer, S. McCarthy, J. Eriksson, T. Kwan, M. Jhamai, K. Trajanoska, Y. Memari, J. Min, J. Huang, P. Danecek, B. Wilmot, R. Li, W. C. Chou, L. E. Mokry, A. Moayyeri, M. Claussnitzer, C. H. Cheng, W. Cheung, C. Medina-Gomez, B. Ge, S. H. Chen, K. Choi, L. Oei, J. Fraser, R. Kraaij, M. A. Hibbs, C. L. Gregson, D. Paquette, A. Hofman, C. Wibom, G. J. Tranah, M. Marshall, B. B. Gardiner, K. Cremin, P. Auer, L. Hsu, S. Ring, J. Y. Tung, G. Thorleifsson, A. W. Enneman, N. M. van Schoor, L. C. de Groot, N. van der Velde, B. Melin, J. P. Kemp, C. Christiansen, A. Sayers, Y. Zhou, S. Calderari, J. van Rooij, C. Carlson, U. Peters, S. Berlivet, J. Dostie, A. G. Uitterlinden, S. R. Williams, C. Farber, D. Grinberg, A. Z. LaCroix, J. Haessler, D. I. Chasman, F. Giulianini, L. M. Rose, P. M. Ridker, J. A. Eisman, T. V. Nguyen, J. R. Center, X. Nogues, N. Garcia-Giralt, L. L. Launer, V. Gudnason, D. Mellstrom, L. Vandenput, N. Amin, C. M. van Duijn, M. K. Karlsson, O. Ljunggren, O. Svensson, G. Hallmans, F. Rousseau, S. Giroux, J. Bussiere,

P. P. Arp, et al. Whole-genome sequencing identifies en1 as a determinant of bone density and fracture. *Nature*, 526(7571):112–7, 2015.

[247] W. Zhu and H. Zhang. Why do we test multiple traits in genetic association studies? *J Korean Stat Soc*, 38(1):1–10, 2009.

[248] Z. Zhu, F. Zhang, H. Hu, A. Bakshi, M. R. Robinson, J. E. Powell, G. W. Montgomery, M. E. Goddard, N. R. Wray, P. M. Visscher, and J. Yang. Integration of summary data from gwas and eqtl studies predicts complex trait gene targets. *Nat Genet*, 48(5):481–7, 2016.

[249] O. Zuk, E. Hechter, S. R. Sunyaev, and E. S. Lander. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc Natl Acad Sci U S A*, 109(4):1193–8, 2012.