Statistical Methods for Large Scale Genetic Analyses

by

Joshua Weinstock

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in the University of Michigan
2021

Doctoral Committee:

Professor Goncalo Abecasis, Chair
Associate Professor Hyun Min Kang
Professor Jun Li
Associate Professor Xiang Zhou

Joshua Weinstock

jweinstk@umich.edu

ORCID iD: 0000-0001-7013-1899

Dedication

I dedicate this dissertation to my wife Ariella, and my parents, Marian and Lew, who provided me much support along the way.

Acknowledgements

I owe a great deal of gratitude to many. In the interest of brevity, I cannot name them all. I would especially like to thank my mentors in the Biostatistics department – Goncalo Abecasis, Hyun Min Kang, Xiang Zhou, and Matthew Zawistowski. It has been a privilege to collaborate with them and learn from their insights. I would also like to acknowledge my fellow students who provided social support and encouragement. Although a dissertation is a challenging personal endeavor, commiserating with friends offered welcome respite.

Table of Contents

List of Tables

List of Figures

Abstract

Population scale genomic analyses have informed the development of novel therapeutics, diagnostics, and understanding of disease etiology. Among the recent developments in human genetic association analyses, electronic health record (EHR) linked biobanks and population scale whole genome sequencing (WGS) have provided fertile ground for association discovery. In tandem with the emergence of these approaches, novel computational and statistical approaches are needed to address the methodological challenges of working with these data.

In Chapter 2, I present study design recommendations and meta-analysis results for genetic association studies applied to clinical laboratory data in EHR linked biobanks. We conducted genome-wide association studies (GWAS) of 70 clinical lab traits from both the Michigan Genomics Initiative (MGI) and BioVU from the University of Vanderbilt health system. In addition to the discovery of novel association results, we conducted systematic study design analyses in parallel across the two biobanks to inform recommendations for association studies of lab traits.

In Chapter 3, I present a novel sparse Mendelian randomization (MR) method for causal inference. MR methods are an instrumental variable approach for inferring the causal effect of an exposure on an outcome using genetic variants as an instrument. Under settings where the proportion of genetic variants that are causal is low, current approaches that assume dense genetic architectures may have poor statistical power. Here, we present a novel Bayesian MR method using a horseshoe prior which can be applied to summary statistics. The horseshoe prior is a continuous-scale shrinkage prior which facilitates variable selection. We use simulations to evaluate the performance of the method across genetic architectures. We apply the method to lab trait GWAS summary statistics.

In Chapter 4, I present a novel method for estimating the rate at which somatic clones are expanding in clonal hematopoiesis. Clonal hematopoiesis refers to a state of mosaicism in blood defined by the acquisition of oncogenic driver mutations at an appreciate clone size and can be identified using WGS. Previous approaches for describing the growth of these mutations have relied on longitudinal sequencing methods. Here, we develop a Bayesian hierarchical model for estimating the parameters that describe the expansion of driver variants. In contrast to previous reports, our method only requires a single draw of blood. We validate the method using simulations and longitudinal amplicon sequencing. We apply our method to ~5,000 samples with clonal hematopoiesis from the Trans-Omics for Precision Medicine (TOPMed) sequencing initiative, enabling association studies of the molecular determinants of clonal expansion.

Chapter 1 Introduction

Background

Analysis of human genetic data has facilitated discovery of disease etiology, novel diagnostics, and therapeutics. Although any two randomly sampled human genomes are highly similar, alleles at a subset of polymorphic sites associate with both disease and non-disease traits. Despite the immense scientific interest in studying genetic variation, until the current millennium, genotyping was very costly. The human genome project cost $2.7 billion USD to complete (Human Genome Project FAQ). Recent advances in genomics have precipitously reduced genotyping and sequencing costs. In 2021, the cost to sequence a whole human genome is approximately $1,000 USD, which is five orders of magnitude smaller than the cost from 2006 (DNA Sequencing Costs: Data). Decreasing costs have facilitated a deluge of human genetics data. In tandem with unprecedented population scale repositories of genotype data, numerous statistical and computational challenges have emerged. Among these, two will be of interest in this dissertation – association discovery in electronic health record (EHR) linked biobanks and variant calling in whole genomes.

Electronic health records are digital containers of patient linked health data. They may include data on disease state, clinical laboratory measurements, free text from physicians, or other sources of data. These data are collected during routine clinical care primarily for billing purposes rather than research use. As EHRs collect numerous variables of interest to biomedical researchers, they represent a potential trove of valuable information. Linking these data to genotypes enables genetic association discovery with the numerous phenotypes.

Causal Inference Using Genetic Instruments and PheWAS

Recently, numerous studies have performed genetic association studies (GWAS) in a comprehensive manner, using every phenotype available in an EHR-linked biobank, an analysis termed PheWAS (phenome-wide association study (Denny *et al.*, 2010)). Much interest has focused on using billing codes to define dichotomous disease phenotypes (Zhou *et al.*, 2018). Similarly, clinical lab traits have emerged as a useful intermediate phenotype for many disease states. Clinical lab traits include lipid measurements, (e.g. low-density lipoprotein), blood cell indices (e.g., white blood cell count), and glucose measurements, among others. Like studying gene expression data, clinical lab traits provide a phenotype that may be upstream of disease and therefore 'closer' in some sense to genetic variation. For disease traits that are defined based on dichotomizing a continuous lab trait, studying the underlying lab may also yield increased power. Parallel efforts to disease PheWAS are needed to facilitate optimal study design and association studies with lab traits. In Chapter 2, I discuss LabWAS (lab-wide association study) and provide study design and methods recommendations for studying genetic variation in this class of EHR derived phenotypes.

LabWAS introduces two primary challenges when compared to the classic PheWAS analysis. First, in contrast to working with ICD10 derived phecodes (Denny *et al.*, 2010; Zhou *et al.*, 2018), clinical lab traits lack a standardized coding scheme. This means that when comparing traits from two different health systems, the same trait may have two different names. The same lab trait may also be measured through different diagnostic assays across health systems, leading to different measurement scales and potential batch effects. As we describe in Chapter 2, the ostensibly facile task of matching lab traits across health systems requires an interdisciplinary team of biostatisticians and pathologists. Second, as clinical lab traits are quantitative traits with informative missingness and longitudinal measurement, different statistical methods than PheWAS are required. The literature provides no consensus on best practices for analysis of clinical lab traits. For example in (Verma *et al.*, 2018), the median summary statistic is used to summarize multiple lab measurements, but in (Kanai *et al.*, 2018), the mean is used and several health state covariates are included in the regressions. In Chapter 2, we comprehensively assess competing statistical procedures for discovery of associated genetic variants, which we perform in parallel across multiple biobanks. We anticipate that our analysis will serve as a reference point for analyses of lab traits in other biobanks.

As GWAS have now been performed across of a variety of phenotypic modalities, researchers have sought to interrogate the causal network among multiple traits. Germline genetic variants have many attractive properties as instruments for causal inference – they are static throughout life, have little measurement error, and in a temporal sense precede the advent of most heritable disease. Crucially, they are frequently independent of environmental confounders. These properties facilitate causal inferences where traditional observational epidemiology may be less permissive of causal interpretation. Mendelian randomization (MR) is an instrumental variable method that uses genetic variation as an instrument for an exposure (Davey Smith and Ebrahim, 2003), for use in inferring the causal effect between an exposure and an outcome phenotype.

With the advent of Biobank derived PheWAS, summary statistics are now abundant, and provide an easy source for genetic instrumentation (Bycroft *et al.*, 2017; Zhou *et al.*, 2018; Hemani *et al.*, 2018). These troves of association summary statistics enable causal inference between two traits that may not even be measured on the same individuals (Pierce and Burgess, 2013). Despite the convenience of MR, appropriate instrument selection remains vexing, a challenge that we further explore in Chapter 3. Instrument selection is in some sense a variable selection problem, and the initial signals from GWAS can be refined through consideration of linkage disequilibrium (LD). In a Bayesian framework, priors that reflect the expected genetic architecture can also provide improved variable selection over standard maximum likelihood (Zhou *et al.*, 2013). In Chapter 3, we explore a novel sparse MR method motivated by these considerations. This method is also partly motivated by the advances in genotyping technologies. As whole genomes have become less costly and genotype imputation methods have become more accessible (Taliun *et al.*, 2021; Das *et al.*, 2016), GWAS have been performed on increasingly larger sets of genetic variants. As genotyping arrays are typically enriched for disease-associated variation, the density of causal variants has likely decreased as variant genotyping expands beyond the sites on genotyping arrays. This suggests that methods that assume sparsity among causal variants are more useful today than the past.

Somatic Variation Derived from Whole Genomes and the Dynamic Human Genome

Most human genetics research has focused on inherited, or germline variation – the differences between human genomes that are inherited from our parents and that remain static throughout life. However, our genomes vary not only between individuals but also within. We are all mosaics of genomes, acquiring mutations as we age (Jaiswal and Ebert, 2019). Beyond the acquisition of mutations, telomeres, which reside at the ends of chromosomes, also decrease with aging (Blackburn, 1991). Most of this acquired variation has no bearing on disease. However, the recent study of somatic variation in blood of healthy adults has indicated that many of us acquire deleterious oncogenic mutations in a small fraction of our blood cells (Jaiswal *et al.*, 2017; Bick *et al.*, 2020). Serendipitously, whole genomes derived from peripheral blood enable interrogation of these dynamic elements without additional assays. However, as population whole genome sequencing is not primarily intended for this purpose, novel computational methods are needed to derive and measure these dynamic quantities.

Large scale WGS enables detection of somatic variation and telomere lengths in blood (Bick *et al.*, 2020; Taub *et al.*, 2020). Somatic variant calls can be used to detect the presence of leukemogenic clonal expansions in blood, a state called clonal hematopoiesis of indeterminate potential (CHIP). CHIP is aging related phenomenon, and previous reports have associated CHIP with increased risk for cardiovascular disease and hematologic malignancy (Jaiswal *et al.*, 2017). CHIP is a heterogenous phenomenon, as it is comprised of a diverse class of mutations and may present in varying fractions of blood cells. The proportion of blood cells harboring a mutation, termed the clone size, has been shown to be predictive of deleterious phenotypic consequences (Bick Alexander G. *et al.*, 2020). This implies that early detection of quickly growing clones may enable earlier therapeutic intervention. However, no methods currently exist for the estimation of clonal expansion from a single draw of blood.

In Chapter 4, we explore the derivation of statistical estimators of clonal expansion using WGS from whole blood. We develop a hierarchical Bayesian model for the estimation of parameters that govern a stochastic process that describes clonal expansion rate. We leverage recent advances in probabilistic programming languages (Stan Development Team, 2020) to efficiently sample from the posterior using Hamiltonian Monte Carlo. Using the largest tranche to date of genomes from CHIP carriers, we also use clonal expansion as a phenotype in GWAS, yielding potential molecular targets for therapeutics that may modulate clonal expansion. We anticipate that these advances will contribute to novel diagnostics and therapeutics for CHIP.

In summary, in this dissertation I make a number of contributions towards modern challenges in computational genetics. I anticipate that our results and methods will further progress in EHR-linked biobanks and population scale repositories of whole genomes.

References

Bick,A.G. *et al.* (2020) Inherited causes of clonal haematopoiesis in 97,691 whole genomes. *Nature*, **586**, 763–768.

Bick Alexander G. *et al.* (2020) Genetic Interleukin 6 Signaling Deficiency Attenuates Cardiovascular Risk in Clonal Hematopoiesis. *Circulation*, **141**, 124–131.

Blackburn,E.H. (1991) Structure and function of telomeres. *Nature*, **350**, 569–573.

Bycroft,C. *et al.* (2017) Genome-wide genetic data on ~500,000 UK Biobank participants. *bioRxiv*, 166298–166298.

Das,S. *et al.* (2016) Next-generation genotype imputation service and methods. *Nature genetics*, **48**, 1284–1287.

Davey Smith,G. and Ebrahim,S. (2003) 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease?*. *International Journal of Epidemiology*, **32**, 1–22.

Denny,J.C. *et al.* (2010) PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics*, **26**, 1205–1210.

DNA Sequencing Costs: Data *Genome.gov*.

Hemani,G. *et al.* (2018) The MR-base platform supports systematic causal inference across the human phenome. *eLife*, **7**.

Human Genome Project FAQ *Genome.gov*.

Jaiswal,S. *et al.* (2017) Clonal Hematopoiesis and Risk of Atherosclerotic Cardiovascular Disease. *New England Journal of Medicine*.

Jaiswal,S. and Ebert,B.L. (2019) Clonal hematopoiesis in human aging and disease. *Science*, **366**.

Kanai,M. *et al.* (2018) Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. *Nature Genetics*, **50**, 390–400.

Pierce,B.L. and Burgess,S. (2013) Efficient Design for Mendelian Randomization Studies: Subsample and 2-Sample Instrumental Variable Estimators. *American Journal of Epidemiology*, **178**, 1177–1184.

Stan Development Team (2020) Stan Modeling Language Users Guide and Reference Manual, 2.17.

Taliun,D. *et al.* (2021) Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature*, **590**, 290–299.

Taub,M.A. *et al.* (2020) Novel genetic determinants of telomere length from a trans-ethnic analysis of 109,122 whole genome sequences in TOPMed. *bioRxiv*, 749010.

Verma,A. *et al.* (2018) PheWAS and Beyond: The Landscape of Associations with Medical Diagnoses and Clinical Measures across 38,662 Individuals from Geisinger. *The American Journal of Human Genetics*, **102**, 592–608.

Zhou,W. *et al.* (2018) Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature Genetics*, **50**, 1335–1341.

Zhou,X. *et al.* (2013) Polygenic Modeling with Bayesian Sparse Linear Mixed Models. *PLOS Genetics*, **9**, e1003264.

Chapter 2 LabWAS: Novel Findings and Study Design Recommendations From a Meta-Analysis of Clinical Labs in Two Independent Biobanks[1]

Introduction

Laboratory testing is a key component of modern medicine. Laboratory measurements provide a glimpse into the functioning of the human body, allowing clinicians to diagnose and monitor disease. In most health systems, lab measurements are routinely captured in patient Electronic Health Records (EHRs) alongside disease diagnoses, free text notes and medical procedures to provide a detailed, longitudinal health history (Carolina and Carolina, 2013). EHRs present exciting research potential by providing broad phenotyping on large cohorts with minimal cost(Wei and Denny, 2015; Hanauer *et al.*, 2015).

Several large-scale genetic studies have already leveraged biobanks linked to EHRs, such as the UK Biobank(Bycroft *et al.*, 2017), Japan Biobank(Nagai *et al.*, 2017), FinnGen(FinnGen) and HUNT(Krokstad *et al.*, 2013), as sources of phenotypes for Genome-wide Association Studies (GWAS) (Bycroft *et al.*, 2017; Nagai *et al.*, 2017; FinnGen; Krokstad *et al.*, 2013)[4–7]. The phenotypes are typically based on International Classification of Diseases (ICD) codes

[1] This chapter has been published in PLOS Genetics. I am a co-first author of the publication.

mapped to dichotomous traits(Wei *et al.*, 2017). Although disease is often thought of in all-or-nothing binary states, many diseases exist on a continuum with the ultimate clinical diagnosis occurring once a relevant quantitative laboratory measurement exceeds a pre-determined threshold. For example, hypercholesteremia, diabetes mellitus and chronic kidney disease are each diagnosed almost entirely on measurements of low density lipoprotein (LDL), glycated hemoglobin (or glucose) and creatinine, respectively. Laboratory measurements can therefore be a more sensitive measure of underlying health than diagnosis and may provide a more powerful outcome for analysis. As an example, the hypercholesterolemia and coronary artery disease risk locus *PSCK9* was initially discovered based on quantitative LDL measurement rather than clinical diagnosis (Chen *et al.*, 2005; Shioji *et al.*, 2004). In contrast to binary disease phenotypes, there are fewer examples of genetic analyses of EHR-derived quantitative lab values(Kanai *et al.*, 2018; Kullo *et al.*, 2010; Klarin *et al.*, 2018). Hereafter, we use the term lab traits to refer to quantitative biomarkers assayed through clinical laboratory testing (e.g., “creatinine", "LDL cholesterol"), and the term lab measurements to refer to realized values of these tests stored in patient EHRs.

The rich source of quantitative lab measurements in EHR cohorts comes with unique concerns. Quantitative traits collected specifically for research purposes typically use a controlled experimental design to ensure consistency among samples. In contrast, lab measurements contained in EHRs are a historical record of medical care. As such, patients may have hundreds of lab measurements for some traits and none for others, depending on their specific health history and utilization of the health system. The measurements can be collected in times of sickness or good health leading to substantial variation in measurements for the same lab. Lab measurements can also be artificially modified by prescription medication, such as statin use for lowering LDL cholesterol. Moreover, recruitment mechanisms and health system demographics can dramatically shape the overall health of the biobank cohort, which in turn dictates lab measurements available for analysis. The broad impact of using such “real world” measurements for genetic association studies is unclear. Questions remain over the effect and robustness of analytic choices made when analyzing EHR-based lab traits including how best to summarize complicated, longitudinal lab measurements and whether comorbid diseases highly correlated with lab measurements must be considered. Prior studies are not consistent in addressing these concerns. For example, GWAS of EHR-derived quantitative traits in Biobank Japan enrolled patients with at least 1 of 47 diagnoses and controlled for all 47 diagnoses while testing each lab (Kanai *et al.*, 2018). In contrast, an analysis of labs within the Geisinger EHR did not control for underlying disease states (Verma *et al.*, 2018). The variety of methods to summarize lab measurements and models to test for genetic association indicates that the question of how to analyze these data remains unsettled.

In this paper we explore strategies for analyzing quantitative lab measurements extracted from EHRs and describe the first large-scale meta-analysis of EHR-derived lab traits across independent health systems. We used lab measurements and genetic data from two academic health systems: the BioVU cohort from Vanderbilt University(Roden *et al.*, 2008) and the Michigan Genomics Initiative (MGI) from Michigan Medicine(Fritsche *et al.*, 2018). Meta-analysis offers a mechanism to increase sample size and power for detecting genetic risk variants but comes with distinct challenges for EHR lab traits, particularly matching lab traits between health systems and determining specific analysis protocols. The cohorts differ dramatically in their recruitment mechanisms, patient composition and recording format for lab measurements:

MGI was predominantly recruited through inpatient surgical encounters at Michigan Medicine whereas BioVU recruitment required outpatient appointments at Vanderbilt University Medical Center. As a result, MGI is enriched for diseases treated surgically such solid tumors (Fritsche *et al.*, 2018). This heterogeneity reflects the reality of EHR-based phenotyping, and strategies must be developed for future collaborative work on the growing number of EHR-linked biobanks.

Our initial challenge was identifying which labs to meta-analyze between the health systems. Accurately matching labs is complicated by the fact that no standardized coding scheme exists for lab measurements. Dichotomous disease traits are readily matched between health systems using the ubiquitous ICD coding system for disease diagnoses(McCarty *et al.*, 2011). Although the Logical Observation Identifiers Names and Codes (LOINC) system offers the promise of interoperability for lab traits, it is cumbersome and maps poorly onto other ontologies(Bodenreider, 2008). For example, there are 21 distinct codes for blood glucose which might not be used consistently between institutions. Moreover, health systems may adopt their own idiosyncratic internal terminology for electronic recording of lab results. Based on a methodical manual review of EHR text descriptions and lab measurements, we identified 70 lab traits between BioVU and MGI that could be matched with high confidence. We extracted previously identified variants for these lab traits from the GWAS catalog to serve as true positive variants for assessing subsequent analyses. Our meta-analysis replicated nearly 75% of these true positive variants, validating both the accuracy of lab matches across health systems and the overall quality of the EHR lab data. Further, we discovered 31 novel lab-associated variants across 22 labs, including the first reported associations for the saliva and pancreatic enzyme amylase and bicarbonate CO2, a gaseous waste product from metabolism carried in the blood.

We immediately replicated 22 (71%) of these novel associations using an independent second set of BioVU samples.

The meta-analysis required several strategic choices regarding data preparation and statistical analysis. We explored the consequences of various analytic choices using a series of mirrored analyses performed in MGI and BioVU. In particular, we varied the summary statistic for lab measurements and the inclusion of covariates to control for comorbid diseases in the GWAS. We compared the results between the independent biobank cohorts to assess consistency of effects. We hypothesized that alternative summary statistics to the basic mean could provide more powerful genetic analyses. We considered: the median lab measurement due to robustness against data recording errors and extreme measurements, the first available lab measurement to mitigate the effects of prescription drugs on modifiable lab traits, and the maximum recorded measurement to magnify variation in extreme measurements. The comorbidity analysis compared GWAS results from models that included indicator covariates for a wide array of diseases to models that did not.

The complete set of GWAS summary statistics from this analysis are broadly available to the research community. We encourage others to use this data to replicate their own GWAS findings and perform hypothesis-driven lookups on specific SNPs or lab traits of interest. Our results are viewable through an interactive *PheWeb* web browser(Gagliano Taliun *et al.*, 2020) at http://pheweb.sph.umich.edu/mgi-biovu-labs and available for bulk download at https://phewascatalog.org/labwas and ftp://share.sph.umich.edu/mgi_biovu_labwas/.

Methods

Datasets

We analyzed data from two university hospital biobanks that link electronic health records with genetic data: BioVU from Vanderbilt University and the Michigan Genomics Initiative (MGI) from Michigan Medicine. We restricted our analysis to unrelated patients of European ancestry because of insufficient patient sample sizes and a paucity of known variants in non-European populations.

The BioVU cohort has been described previously(Roden *et al.*, 2008). Briefly, DNA was extracted from surplus blood samples and genotyping data was linked to de-identified EHR data. For this study, we used a cohort of 20,515 individuals genotyped on the Multi-Ethnic Genotyping Array (MEGA) from Illumina and estimated to be of European ancestry by admixture(Alexander *et al.*, 2009). We included 843,242 SNPs that passed standard marker QC filters and had a minor allele frequency >1%. We retrieved all available lab measurements in this cohort that occurred when the subject was at least 18 years of age.

The MGI cohort has also been described previously(Fritsche *et al.*, 2018). Briefly, MGI samples were recruited primarily through surgical encounters at Michigan Medicine and provided consent for linking of their EHRs and genetic data for research purposes. MGI samples were genotyped on customized Illumina HumanCoreExome v12.1 bead arrays. European samples were identified using Principal Component Analysis. We used a data freeze consisting of 37,354 unrelated European individuals for this analysis. MGI samples were imputed to the Haplotype Reference Consortium using the Michigan Imputation Server(Das *et al.*, 2016), providing ~14 million SNPs with a minimac imputation quality R2>0.3 and an allele frequency greater than 1e-6. We analyzed the set of ~800K overlapping SNPs between the MGI imputed genotypes and the BioVU MEGA array for this study.

Harmonization of Labs Between Health Systems and the GWAS Catalog

We extracted all available clinical lab measurements and metadata from the electronic health records of MGI samples and BioVU samples. `We collapsed distinct labs when obvious duplications were present (e.g., "Eosinophils" and "EOSINOPHILS"). Available metadata differed slightly between the health systems but included brief text descriptions, unit of measurements, and range for normal values. We excluded individual lab measurements taken outside the health system labelled as "External." In cases where multiple tests examined the same analyte, e.g. blood glucose, we removed point of care (POC) tests which are more susceptible to technical artifacts and tend to be deployed in intensive care or emergency settings where acute disease or treatment effects supervene determinants of the underlying baseline (Nichols, 2011; Larsson *et al.*, 2015). Lab traits were matched between the Vanderbilt and Michigan health systems based on manual curation of the metadata including recorded lab names, clinical descriptions, measurement units, range of measurements, and patient count.

Disease phenotypes

In order to study the effect of underlying health conditions we extracted ICD9 and ICD10 diagnosis codes from the EHR of the BioVU and MGI cohorts. We searched for diagnosis for 42 diseases with the potential to alter a clinical lab measurement (Supplementary Table). We started with the disease list used in the BioBank Japan lab analysis(Kanai *et al.*, 2018) and removed diseases which do not occur in our population (e.g. febrile seizures of infancy) and those expected to have minimal effect on labs (e.g. cataracts). We supplemented their list with chronic diseases expected to have a large impact on labs due to their prevalence (e.g. hypertension). We created an indicator variable for each disease (1 if the sample had at least one qualifying ICD code for the specific disease and a 0 otherwise) to include as covariates in GWAS regression analyses.

Statistical Analysis

Intra-cohort Genome-wide Association Studies

We first performed GWAS analysis of each lab trait separately in the MGI and BioVU cohorts. We performed multiple GWAS for each lab, varying the statistic used to summarize the longitudinal lab measurements for each sample (mean, median, first available measurement and maximum available measurement) and the inclusion of binary indicators for diagnosis comorbid diseases in the GWAS regression.

For each GWAS, the distribution of lab summary statistics was inverse normalized separately within the MGI and BioVU cohorts prior to regression analysis. In a separate analysis of the BioVU cohort, we determined that inverse normalization of lab values performed better than applying no transformation, or a log or square root transformation for controlling GWAS type I error. Genome-wide association tests were performed on the inverse normalized traits using additive linear regression models containing age, sex and four principal components as covariates. The comorbidity model controlled for disease status by inclusion of an additional 42 covariates for the binary disease phenotypes. The regression analyses were performed in the BioVU cohort using PLINK(Purcell *et al.*, 2007) and in the MGI cohort using *epacts 3.3.0* (Kang, 2014).

Comparison of p-values across cohorts

We treated the GWAS of mean trait value with no disease covariates as the default. We quantified the impact of each alternate analysis strategy relative to the default analysis by computing the log fold change in p-value between the alternative and default analysis for each analyzed SNP. That is, for each SNP we compute the quantity

$$\Delta_p = -\log_{10}(p\text{-}value\ for\ alternative\ analysis\ /\ p\text{-}value\ for\ default\ analysis\ )$$

for the MGI analysis and the BioVU analysis separately. A positive value of Δ_p indicates a SNP that increases in significance (smaller p-value) when the alternate summary statistic. A negative value of Δ_p indicates a decrease in significance for the alternate analysis. Scatterplots of Δ_p computed in MGI and BioVU summarize the magnitude and consistency of change in p-value significance between the cohorts (Figure 4 for an example, Supplementary Material). We performed LD-pruning on non-catalog SNPs to simplify the scatterplots Since most SNPs are not associated with the lab trait of interest, alternative summarizations simply result in independent noise between the two cohorts, resulting in a diamond shaped pattern centered at the origin. We implemented a heuristic to formally distinguish the SNPs with largest changes in p-value between the alternative and default analysis methods from those with movement due simply to random noise. The heuristic generates a bounding quadrilateral polygon around the diamond cluster of points, generated using simulated annealing to determine the bounding coordinates of a polygon containing 99.9% of all SNPs. We defined SNPs outside the boundaries of the polygon as those with largest simultaneous changes in p-values in both cohorts. Catalog SNPs located outside the bounding polygon were classified as having either a concordant increased effect if p-value significance increased in both MGI and BioVU, a concordant decrease effect if p-value significance decreased in both MGI and BioVU or a discordant effect if the p-value increased in significance in one cohort but decreased in the other.

Meta-analysis

We meta-analyzed the GWAS results from the MGI and BioVU default analysis (mean trait value, no disease covariates). The meta-analysis was performed using *METAL* by combining study-specific GWAS effect size estimates and standard errors (Willer *et al.*, 2010). We computed genomic control inflation factors (λ_{GC}) on a set of LD-pruned SNPs for each meta-analyzed lab.

GWAS Catalog Variants

We created a list of previously identified genetic associations for each analyzed lab trait using the GWAS catalog(Buniello *et al.*, 2019) (downloaded 9/27/2017). We searched the catalog for quantitative phenotypes matching our analyzed labs using pattern matching in the DISEASE_TRAIT, MAPPED_TRAIT, and P_VALUE_TEXT columns. We searched for each lab using multiple potential string patterns, for example "AST", "aspartate aminotransferase", "SGOT", and "serum glutamine oxaloacetic aminotransferase". For purposes of replication, we limited our catalog search to studies of European cohorts performed on adults of both sexes without disease-based sampling (e.g. glucose measurements in type 2 diabetes samples) and required a reported p-value of at least 5e-8. We considered a catalog association replicated if the meta-analysis p-value for our corresponding lab was < 0.05 and the BioVU and MGI studies had the same direction of effect.

Definition of novelty

We report novel lab-SNP associations as those reaching genome-wide significance that have not been previously reported in European populations and are not reasonably expected based on existing SNP-lab associations in similar labs. We used the following criteria: meta-analysis p-value <5e-8, consistent direction of effect between MGI and BioVU and at least 1 megabase from any previously reported SNP for the given lab or a related lab in the GWAS catalog. Here, we define related labs as those which are commonly ordered as part of a panel of correlated tests (e.g. AST and ALT for liver function) or arithmetically-dependent traits (e.g. LDL and total cholesterol), and therefore likely to indicate the same biological association. We report the "peak" or most significant SNP when a group of novel SNPs are in linkage disequilibrium.

Replication of Novel Associations

We performed a replication analysis of novel associations identified in the meta-analysis using an independent cohort of BioVU samples that became available after the original meta-analysis was performed. This replication cohort consisted of 29,043 European ancestry adult individuals with extant lab data recruited using the same procedure as the initial BioVU cohort, genotyped on the same MEGA genotyping array, and subjected to the same data QC procedure. We declared a novel association to be replicated if the replication p-value was <0.05 and the direction of effect was consistent with that from the meta-analysis.

Ethics statement

Data were collected according to Declaration of Helsinki principles. MGI study participants' consent forms and protocols were reviewed and approved by the University of Michigan Medical School Institutional Review Board (IRB ID HUM00099605 and

HUM00155849). Opt-in written informed consent was obtained for each MGI participant. BioVU is Vanderbilt University's biobank of DNA extracted from leftover and otherwise discarded clinical blood specimens. BioVU operates as a consented biorepository; all individuals must sign the BioVU consent form in order to donate future specimens.

Results

We extracted all available clinical lab measurements from the electronic health records (EHRs) for genotyped samples in two academic biobank cohorts: the Michigan Genomics Initiative(Fritsche *et al.*, 2018) (MGI) at Michigan Medicine and the BioVU(Roden *et al.*, 2008) at Vanderbilt University. In total, this consisted of 35,785,074 lab measurements in 50,743 MGI samples, and 28,929,660 lab measurements in 61,378 BioVU samples. We focused on samples of European ancestry in both cohorts due to insufficient sample sizes in other ancestry groups. Genetic analyses were performed on the set of ~800K overlapping SNPs between the MGI imputed genotypes and the BioVU MEGA array genotypes.

We analyzed 70 labs matched with high confidence between the health systems and having at least 1,000 samples with the lab measured in each health system by(Table 1). We searched the GWAS catalog for known genetic associations among the 70 lab traits to serve as "true positive" variants to validate the data and assess competing analysis strategies. We identified 4,140 such associations, of which, 1,313 (32%) across 48 different traits were in the set of overlapping markers tested in the meta-analysis . Many lab traits have been well studied (Willer *et al.*, 2013; Astle *et al.*, 2016) and provided many testable catalog SNPs. LDL, for example, had 84 catalog SNPs that could be directly tested in our meta-analysis. Alternatively, several labs had relatively few or no catalog SNPs, including labs for which either no variant was reported in the catalog or the catalog variants were not typed in at least one of our cohorts.

Meta-Analysis of Labs in MGI and BioVU

The 70 EHR-derived lab traits were first analyzed separately in the cohorts using the mean measurement as the individual-level outcome. The meta-analysis sample size differed between labs, ranging from 7,429 for uric acid to 46,382 for hematocrit (Figure 1), reflecting the frequency with which different labs are administered in health systems. Several labs have previously been studied in much larger cohorts, including the differential panel of 10 white blood cell measures, analyzed in >170K samples in the UK BioBank(Astle *et al.*, 2016). However, this meta-analysis provides the largest sample size for 34 labs, including 14 clinical lab traits with no previously reported study in the GWAS catalog at the time of our analysis. Genomic control lambda values (λ_{GC}) confirmed the meta-analyses were well-controlled(Devlin *et al.*, 2001). The mean λ_{GC} across all labs was 1.035, ranging between 0.995 and 1.103. Consistent with polygenicity(B. K. Bulik-Sullivan *et al.*, 2015), traits with a larger numbers of catalog variants had, on average, larger λ_{GC} values. The mean λ_{GC} for labs with zero testable catalog SNPs was 1.020. Labs with one to twenty testable Catalog SNPs had mean λ_{GC} of 1.028 and labs with greater than 20 testable Catalog SNPs had mean λ_{GC} of 1.066.

Replication of GWAS Catalog SNPs

We first performed a replication analysis of the 1,313 GWAS catalog SNPs to validate the EHR-derived lab traits. Overall, we replicated 982 of the GWAS catalog SNPs, giving an overall replication rate of 74.8% (Table 1). Replication rates varied across the individual labs;

however, we did replicate at least one catalog SNP for each of the 48 traits with a testable catalog SNP. Replication rates were high for several previously well-studied traits, including red blood cell indices (MCHC, MCH, MCV), metabolic measures (glucose and HgbA1C) and creatinine. The lowest replication rates occurred for the differential panel of white blood cell traits (neutrophils, lymphocytes) which included catalog SNPs discovered in the much larger UK Biobank cohort(Bycroft *et al.*, 2017). Interestingly, replication rates differed among the well-studied lipid panel traits. We replicated a lower percentage of catalog SNPs for LDL cholesterol and total cholesterol compared to triglycerides and HDL cholesterol.

Several factors influenced our ability to replicate individual catalog SNPs (Figure 2), each consistent with statistical power rather than adequate matching of labs as the primary limiting factor. Replication increased sharply with the number of publications reporting the association, as quantified using the PMID citation count from the GWAS catalog (Figure 2A). Associations reported only once in the catalog are a mix of true unreplicated associations and false positives, whereas associations reported more than once have already been replicated and are likely real. We replicated 70% (699 of 1000) of associations reported only a single time. That rate increased to 77% (196 of 256) for associations reported twice, 91% for associations reported three times and nearly 100% (56 of 57) for associations reported four or more times. Importantly, this analysis provides the first replication for 699 previously reported quantitative lab trait associations, increasing the likelihood that these are true genotype-phenotype associations (Supplementary Table).

Replication rate was also dependent on both the best previously reported p-value for the association and the sample size of the study reporting the association (Figure 2B & 2C). Our replication rate was lowest, between 55%-65%, for associations whose best reported p-value was just above genome-wide significance of 5e-8 but increased sharply thereafter. We replicated ~85% of catalog SNPs with best reported p-value <1e-15 and over 90% of catalog SNPs with best p-value <1e-20. Replication rate increased with the relative size of our meta-analysis compared to the largest reported study. We replicated approximately 90% of catalog SNPs for which our meta-analysis was at least as large as prior studies reporting the association.

Novel Associations

We identified 264 SNP-lab trait pairs representing potentially novel associations. Based on visual inspection, these SNPs corresponded to 31 distinct peaks for which we report the lead SNP having the strongest association signal at each peak (Table 2).

We performed a replication analysis of the 31 lead SNPs using an independent cohort of 29,043 BioVU patients that became available after the initiation of our primary analysis. . One SNP potentially novel for both immature granulocytes measures failed QC filtering in the replication cohort and could not be tested for replication. In total, we replicated 22 of the 31 (71%) novel associations (Table 2). Among the 24 replicated novel SNPs are the first associations for amylase (Amyl) and bicarbonate (CO2). We identified and replicated additional associations for alanine aminotransferase (ALT), alkaline phosphate (AlkP), Relative count of basophils (BasoR), total bilirubin (Bili), calcium (Ca), creatinine phosphokinase (CPK), glucose (gluc), mean corpuscular hemoglobin concentration (MCHC), lipase, and thyroid stimulating hormone (TSH).

Several of our novel findings have biological or existing evidence that support the association. Three of the associations have recently been identified for the same lab in non-European cohorts. rs855791, a missense variant in *TMPRSS6* (transmembrane serine protease 6), and rs8022180, an intronic variant in *TRAF3*, were shown to be associated with bilirubin and serum total protein level, respectively, in a Japanese population(Kanai *et al.*, 2018). rs112574791 is in the glutamic--pyruvic transaminase gene *GPT*, a gene associated with alanine aminotransferase levels in the Korea Biobank(Moon *et al.*, 2019). Our results confirm these prior findings and suggest a cross-ethnic effect in European populations.

The intronic variant rs8051363 in *CTRB1* was associated with both amylase and lipase, clinical assays of pancreas function used to diagnose pancreatitis. While the SNP itself has previously been linked to blood protein measurements(Sun *et al.*, 2018), the *CTRB1* gene encodes chymotrypsin, a component of digestive enzyme secreted by the pancreas, and was previously shown to be associated with alcoholic chronic pancreatitis(Rosendahl *et al.*, 2018). A second novel SNP for lipase, rs9377343, is an intronic variant in *FUT9*, a gene that showed association with diabetic neuropathy in a trans-ethnic meta-analysis(Iyengar *et al.*, 2015).

The amylase-associated SNP rs1930212 resides near three amylase genes (*AMY2B*, *AMY2A* and *AMY1*) on chromosome 1, each of which encodes enzymes that digest starch into sugar(Usher *et al.*, 2015). Copy number variation for amylase genes is hypothesized to have been subject to selective sweeps corresponding to starch content in human diets(Inchley *et al.*, 2016). The rs1930212 SNP tags a known deletion of *AMY2A*, a pancreatic amylase enzyme, most common in populations historically lacking starch rich diets(Inchley *et al.*, 2016).

One of our novel results for calcium, rs2839899, is an intronic variant in *GNAQ* (G protein subunit alpha q), a signaling protein involved in response to various hormones. Variation in *GNAQ* is associated with Sturge-Weber syndrome(Shirley *et al.*, 2013), a hereditary vascular malformation syndrome which can lead to deposits of calcium (calcification) in the brain.

Three SNPs showed associations with glucose. rs7607980 is a missense variant in *COBLL1* previously linked to fasting blood insulin and Type 2 diabetes(Kooner *et al.*, 2011; DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium *et al.*, 2014; Morris *et al.*, 2012). rs9273364 is located near HLA-DQB1-AS1, a gene associated with T2D(Xue *et al.*, 2018). And, although it did not replicate in our analysis, rs896854, a variant mapping to both *NDUFAF6* and *TP53INP1,* has recent associations with T2D(Voight *et al.*, 2010) and eosinophil count(Kichaev *et al.*, 2019) among UK biobank participants.

We note that several associations occurred within the HLA region on chromosome 6, notably for glucose, hemoglobin A1C, and TSH. These variants are likely segregating with HLA types, which are strongly associated with various autoimmune diseases including diabetes and autoimmune thyroiditis, which have strong effects in these particular labs.

Genetic Correlation of Clinical Labs

We computed the genetic correlation between pairs of labs using LD score regression(B. Bulik-Sullivan *et al.*, 2015) to learn about shared genetic basis of these traits (Figure 3). We restricted analysis to the 50 lab traits with heritability of at least 7% based on recommendations by the developers of LDscore regression that estimation of genetic correlation can be unreliable when one of trait has heritability close to zero. We observe strong positive correlations among lab traits of similar function. For example, the liver enzymes alanine aminotransferase (ALT) and aspartate aminotransferase (AST) were strongly correlated, as were the measures of renal function Blood Urea Nitrogen (BUN) and creatinine (Creat). Prothrombin time (PT), a measure of clot formation time and a derivative measure International Normalized Ratio (INR) were positively correlated as expected. More surprisingly, INR was also positively correlated with

vitamin D. While vitamin K is known to be required for the formation of prothrombin, the correlation with Vitamin D suggests a potential covariance in nutrition or nutrient absorption.
A prominent cluster of labs (top right corner of the heatmap) contains primarily white blood cell traits including measures of immature granulocytes, lymphocytes, monocytes and neutrophils. The immature granulocytes also showed a strong correlation with ferritin (ferrit), an iron storage and acute phase protein. Ferritin and immature granulocytes can both be elevated during severe acute inflammation, explaining this correlation.

As expected, HgbA1C and glucose were strongly correlated. More interestingly, they also clustered with Red cell Distribution Width (RDW) and Erythrocyte Sedimentation Rate (SedRat). This cluster of labs showed negative associations with high density lipoprotein (HDL), mean cell hemoglobin concentration (MCHC), and mean cell hemoglobin (MCH). This supports a pathophysiology where the metabolic syndrome (obesity, elevated glucose, low HDL) is linked by complex mechanisms to persistent low-level inflammation (elevated SedRat), and anemia of chronic disease (elevated RDW, low MCH, low MCHC).

We identified a cluster containing red cell indices – mean cell hemoglobin concentration (MCHC), mean cell hemoglobin (MCH), and mean cell volume (MCV) – with total bilirubin (Bili) and transferrin saturation (%SAT). This reflects the biology of hemoglobin – iron is carried to red cell precursors by transferrin and incorporated into heme and thence hemoglobin, red cells are filled with hemoglobin, and at the end of a red cell lifecycle, heme is broken down into bilirubin.

Additional clusters include (1) calcium (Ca), albumin (Alb) and total protein in blood (TProt), (2) thyroid stimulating hormone (TSH) and lactate dehydrogenase (LDH), and (3) hematocrit (HCT), red blood cell count (RBC) and hemoglobin (Hgb) with free tetraiodothyronine (FT4).

Albumin (Alb) is the major blood protein, so Alb levels are unsurprisingly correlated with total blood protein (TProt). Calcium homeostasis is driven by free calcium, while albumin acts as a calcium sink, therefore calcium (Ca) levels would reasonably be expected to correlate with Alb (Payne *et al.*, 1973).

Hematocrit (HCT), red blood cell count (RBC) and hemoglobin (Hgb) are interrelated measures of oxygen carrying capacity in blood and unsurprisingly correlated. In our study, they are also correlated with free tetraiodothyronine (FT4). Anemia (low HCT, RBC, and Hgb) may be a feature of hypothyroidism (low FT4), and tetraiodothyronine - thyroid hormone - has been reported to play a role in red cell maturation (McDermott, 2009; Gao *et al.*, 2017).

A final cluster was identified linking thyroid stimulating hormone (TSH) to lactate dehydrogenase (LDH). Muscle breakdown, manifesting as weakness, is a feature of hypothyroidism, and therefore other laboratory anomalies seen in hypothyroidism include release of muscle enzymes including LDH (McDermott, 2009; Chertow *et al.*, 1974).

Analytic strategies for EHR-derived lab traits

We explored the impact of analytic choices on downstream analysis by performing parallel GWAS analyses in the MGI and BioVU cohorts with one of the analytic steps perturbed from our original analysis: either the individual-level statistic used to summarize longitudinal lab measurements (median, maximum measurement, first available measurement) or the inclusion of covariates for underlying comorbid health conditions. We performed these analyses on the 22 lab traits for which there were least 20 testable GWAS catalog SNPs, using the catalog SNPs to interpret the effect of each analytic strategy on true risk variants.

Summary statistic

Overall, 13.3% of testable catalog SNPs showed a major change in significance when using the median as opposed to mean value for the summary statistic (Table 3). The median rarely resulted in a consistent improvement for both MGI and BioVU. Only 0.4% of catalog SNPs had concordant increased effect compared to 7.6% with concordant decreasing effect and 5.2% with a discordant effect. Creatinine was the sole lab for which using median lab value had a greater number of catalog SNPs with concordant increased significance than catalog SNPs with concordant decreased significance. Even here the effect was small, only two of the 36 catalog SNPs had a concordant increase in significance.

In comparison, the first available measurement and the maximum measurement had a greater impact on association p-values for catalog SNPs. In both cases, the alternate summary statistic was most likely to cause a concordant decrease in significance. Using the first available measurement resulted in concordant increase for only 3.1% of catalog SNPs, whereas 16.9% of catalog SNPs had a concordant decrease and 4.5% had discordant changes in significance. Using the maximum available measure had similar performance (5.6% concordant increase, 18.3% concordant decrease, 5.5% discordant).

Despite an overall trend of reducing significance of known risk variants, several related labs for blood oxygen carrying capacity did benefit from using the first available or maximum measurements. Red blood cell count (RBC), hematocrit (HCT) and hemoglobin (Hgb) each showed concordant increase in significance for several of their respective catalog SNPs without negatively impacting remaining catalog SNPs. This likely reflects red cell biology. Conditions that decrease oxygen carrying capacity, such as blood loss or iron deficiency are far more common than those that increase it, polycythemia vera or severe obstructive sleep apnea, for example. Thus, maximum measurement of an individual's oxygen carrying capacity more likely represents the genetically determined set point.

Controlling for comorbid disease

The comorbidity model, containing binary covariates for 42 comorbid diseases with the potential to alter lab values, produced the largest proportion of catalog SNPs (6.2%) with concordant increased significance in MGI and BioVU among the alternate analysis strategies considered. Despite this, a roughly equal number of catalog SNPs had discordant effects (6.8%) between the two cohorts.

The clearest example of a substantial and consistent effect on catalog SNPs between MGI and BioVU was for HDL and Mean platelet volume (MPV). Interestingly, in contrast to this result for HDL , LDL had no catalog SNPs with concordant increase in significance and seven catalog SNPs with concordant decrease.

Discussion

This study represents the first cross-health system study of EHR-derived lab traits at large scale. We performed meta-analysis GWAS of 70 lab traits and have made these association results easily accessible to the research community. Thoroughly dissecting each lab-SNP combination is a daunting task. Here, we focused on replication of GWAS catalog variants to validate our data and highlighted novel genetic associations. We anticipate that our full results, including those which do not reach genome-wide significance will be useful in replicating future novel results, in studies which synthesize findings across multiple SNPs, or in hypothesis-driven studies which require less stringent thresholds.

Our study serves as a proof-of-principle for performing cross-health-system genetic analysis of EHR-derived lab traits. The high replication rate for known GWAS variants indicates that EHR lab traits can be well-matched between discordant health systems and that measurements taken during real-life medical interactions sufficiently reflect those taken under more idealized experimental conditions. Moreover, this implies that mechanisms underlying variation in lab traits among healthy populations also act in a health system population with diseased individuals, strengthening their clinical relevance. By comparing various analytic strategies, we show that there is no optimal strategy that holds across all lab traits. In fact, we observed many instances in which the alternate analysis simultaneously increased significance for some risk variants and decreased significance for others. Thus, even within a given lab trait, an optimal strategy for variant discovery might not exist. We also considered a summary statistic based on Area Under the Curve for the longitudinal lab data (Tai, 1994; Wolever and Jenkins, 1986) (citation). Analysis in the MGI cohort showed that this measure performed consistently worse than the mean lab measurement (Supplementary Material). A potential area of future research would be determining if multiple versions of a lab trait can be combined into an omnibus test that simultaneously increases power across all risk variants. We encourage researchers to use our results across the various analysis strategies to guide decisions about how best to analyze their traits of interest.

The primary strength of our study was the access to two independent biobank cohorts. Using two cohorts provides an increase in sample size and power over analyzing and reporting on each cohort separately. In addition, the two-cohort design adds a built-in internal consistency check to our results by requiring effect sizes to be in the same direction in both cohorts. This additional requirement reduced the potential for unknown biases in the health system cohorts to create spurious results when replicating GWAS catalog SNPs or novel association discovery. Further, the independent cohorts provided the means to rigorously examine the portability of analytic strategies between biobanks. A similar analysis performed in a single cohort could produce recommendations over fitted to one specific context. Use of multiple sites increases the generalizability of our recommendations. This study was further strengthened by the fortuitous availability of an independent tranche of BioVU samples that provided an immediate replication cohort for the novel findings of our meta-analysis.

Our study has implications for the design and analysis of similar studies in the future. Matching and analyzing lab data between health systems is difficult and requires substantial content knowledge. This study benefited from a multi-disciplinary team consisting of clinical experts to lead the categorization of the raw lab data extracts and statistical geneticists to guide analytic strategies. We leaned heavily on GWAS catalog SNPs to serve as positive controls. We recommend researchers to incorporate an explicit replication step to validate lab data prior to testing novel hypotheses. Summarizing the longitudinal measurements simply using the mean proved relatively robust across labs but was by no means optimal in all scenarios. Future studies can benefit from considering a summary statistic suited to the specific lab trait being evaluated. Our analysis also highlights that close attention must be paid to differences in the preparation and analysis of EHR phenotypes, particularly longitudinal lab measurements. Failing to replicate a prior finding can be due to lack of a true effect but also a variety of differences between biobank cohorts and analytic procedures.

We were motivated to examine the effect of controlling for disease status because of its use in the analysis of lab traits in BioBank Japan(Kanai *et al.*, 2018). Controlling for diseases or risk factors such as tobacco use is a common practice(Astle *et al.*, 2016). We considered testing the effect of each disease individually but discarded it as cumbersome. Our strategy reflects a broad-spectrum approach in which diagnoses that are rare or have limited effect on a lab can be rationalized as not causing harm by remaining in the model. The effect of controlling for comorbid diseases can be unpredictable. For example, within the components of a lipid panel, controlling for disease status led to a net improvement for HDL catalog SNPs, a net worsening for LDL catalog SNPs, and had cohort-specific impact on triglycerides. From a methodological standpoint, this argues for careful consideration of comorbid disease covariates. From a practical standpoint, the absence of diagnostic data should not be seen as precluding use of a clinical lab data.

A limitation of studying clinical labs in real-life cohorts is that some measurements will be affected by medication. We were unable to formally address the effect of medication because of unreliable measurements of medication. However, it remains an important consideration for future EHR-based lab studies and requires further study. There was indication that in situations where a disease diagnosis is likely to be accompanied by medication, for example a diagnosis of dyslipidemia with lipid labs, controlling for disease status diagnosis serves as a reasonable proxy to treatment status. As research interest in EHR phenotypes increases, we anticipate that improved capture of prescription data will facilitate the identification of medication effects.

A further limitation of this study is the number of analyzed genetic variants. The study was restricted to ~800K SNPs because BioVU imputed genotypes were unavailable at time of analysis. Although this limited our ability to discover novel variation, the number of SNPs was more than sufficient to perform the primary purpose of the paper, a proof-of-principle replication analysis across a broad range of clinical labs and analytic strategies. However, there are likely many loci remaining to be discovered for these labs, particularly the understudied traits.

In conclusion, we report the first lab-wide genome-wide association study linking data between two independent EHR-based cohorts. We achieved a high degree of replication of prior associations and report a modest number of new associations. In melding these data sets, we addressed key questions in design and analysis of 'real world' data that are increasingly relevant.

Acknowledgements

The authors acknowledge the University of Michigan Precision Health Initiative and Medical School Central Biorepository for providing biospecimen storage, management, processing and distribution services and the Center for Statistical Genetics in the Department of Biostatistics at the School of Public Health for genotype data curation, imputation, and management in support of this research.

We thank Sebastian Zöllner for valuable feedback on initial drafts.

Analysis on BioVU samples was supported by the Beyond PheWAS R01 grant#: R01 LM010685.

Finally, we thank the three anonymous reviewers whose comments improved the quality of the manuscript.

Table 2.1: Summary of clinical lab traits tested, including meta-analysis samples size, number of testable GWAS catalog SNPs, number of replicated catalog SNPs and replication rate

Lab Name	Category	Description	Meta-Analysis Sample Size	Number of Testable GWAS Catalog SNPs	Number of Catalog SNPs Replicated in Meta-Analysis	Replication Rate (%)
Alb	Liver function	Albumin, most abundant blood protein	39,513	5	4	80
AlkP	Liver function	Alkaline phosphatase, bile duct and bone enzyme released by damage	39,809	3	1	33
ALT	Liver function	ALanine aminoTransferase, liver enzyme released by damage	40,116	0	0	N/A
Amyl	Pancreas	Amylase, digestive pancreas enzyme released by damage	10,368	0	0	N/A
AST	Liver function	ASpartate aminoTransferase, liver enzyme released by damage	40,176	0	0	N/A
BasoAB	Differential	Basophils, white blood cell type (absolute number)	29,653	19	12	63
BasoRE	Differential	Basophils, white blood cell type (relative proportion)	32,578	11	7	64
BEAR	Blood gas	Base Excess ARterial, Acid-base measure of metabolic acidosis or alkalosis	8,895	0	0	N/A
Bili	Liver function	Total Bilirubin, heme byproduct excreted by liver	38,416	4	4	100

BNP	Heart failure	Brain Natriuretic Protein, Signaling protein from heart under stress	9,369	1	1	100
BUN	Renal function	Blood Urea Nitrogen Protein byproduct excreted by kidneys	45,922	0	0	N/A
Ca	Electrolytes	Calcium, blood electrolyte	46,100	9	7	78
Chol	Lipid panel	Total cholesterol	23,642	91	60	66
CKMBRe	Cardiac markers	Creatine Kinase Muscle Brain isoform, relative, Enzyme in heart released by damage	10,964	0	0	N/A
Cl	Electrolytes	Chloride, blood electrolye	45,920	0	0	N/A
CPK	Cardiac markers	Creatine PhosphoKinase, enzyme in skeletal and cardiac muscle released by damage	15,150	0	0	N/A
Creat	Renal function	Creatinine, creatine byproduct excreted by kidneys	46,027	36	29	81
CRP	Inflammatory	C-reactive protein, marker of inflammation	12,447	16	7	44
EoAB	Differential	Eosinophils, white blood cell type (absolute count)	29,912	31	25	81
EoRE	Differential	Eosinophils, white blood cell type (relative proportion)	26,980	28	18	64
Ferrit	Iron	Ferritin, iron storage protein	11,744	6	1	17
FT4	Thyroid function	Free tetraiodothyronin, active thyroid hormone	15,868	0	0	N/A
Gluc	Metabolic	Blood glucose	46,027	18	16	89
HCO3 (CO2)	Blood gas	Bicarbonate, main blood pH buffer	45,932	0	0	N/A

HCT	Complete blood count	Hematocrit, measure of blood oxygen carrying capacity	46382	36	20	56
HDL	Lipid panel	High density lipoprotein cholesterol	23,318	101	84	83
Hgb	Complete blood count	Hemoglobin, oxygen carrying protein	46,159	34	18	53
HgbA1C	Metabolic	Hemoglobin A1C, measure of blood glucose over previous 90 days	17,407	11	10	91
IGranAB	Differential	Immature granulocytes, immature white blood cell type (absolute count)	30,744	0	0	N/A
IGranRE	Differential	Immature granulocytes, immature white blood cell type (relative proportion)	30,683	0	0	N/A
INR	Coagulation	International Normalized Ratio, derivative of PT used to dose anticoagulants	33,695	0	0	N/A
Iron	Iron	Iron	11,317	4	3	75
K	Electrolytes	Potassium, blood electrolyte	45,941	0	0	N/A
LAC	Blood gas	Lactic acid, marker of tissue hypoxia	8,792	0	0	N/A
LDH	Tumor markers	Lactate dehydrogenase, enzyme found in many cell types released by damage	9,734	0	0	N/A
LDL	Lipid panel	Low density lipoprotein cholesterol	22,896	84	58	69
Lipase	Pancreas	Lipase, digestive pancreas enzyme released by damage	12,649	2	2	100

LymphAB	Differential	Lymphocytes, white blood cell type (absolute count)	32,548	35	22	63
LymphRE	Differential	Lymphocytes, white blood cell type (relative proportion)	32,553	20	10	50
MCH	Red cell indices	Mean corpuscular hemoglobin, used to differentiate causes of anemia	46,159	64	57	89
MCHC	Red cell indices	Mean corpuscular hemoglobin concentration, used to differentiate causes of anemia	46,157	20	19	95
MCV	Red cell indices	Mean corupuscular volume, used to differentiate causes of anemia	46,153	77	68	88
Mg	Electrolytes	Magnesium, blood electrolyte	22,773	4	4	100
MonoAB	Differential	Monocytes, white blood cell type (absolute count)	32,587	43	32	74
MonoRE	Differential	Monocytes, white blood cell type (relative proportion)	32,594	15	12	80
MPV	Coagulation	Mean platelet volume	40,058	84	73	87
Na	Electrolytes	Sodium, blood electrolyte	45,933	0	0	N/A
pCO2	Blood gas	Arterial partial pressure of CO2, measure of ventilation	9,516	0	0	N/A
pH	Blood gas	Arterial pH	10,279	0	0	N/A
Phos	Electrolyte	Phosphorus, blood electrolyte	21,618	5	4	80
PLT	Complete blood count	Platelet count, clot forming measure	46,145	102	84	82

PMNAB	Differential	Neutrophils, white blood cell type (absolute count)	32,595	35	15	43
PMNRE	Differential	Neutrophils, white blood cell type (relative proportion)	29,435	21	7	33
pO2	Blood gas	Arterial partial pressure of oxygen, measure of oxygenation	9,557	0	0	N/A
PT	Coagulation panel	Prothrombin time, clot forming measure	33,671	1	1	100
PTT	Coagulation panel	Partial Thromboplastin Time, clot forming measure	30,972	9	6	67
RBC	Complete blood count	Red Blood Cell count, measure of blood oxygen carrying capacity	46,158	50	31	62
RDW	Red cell indices	Red cell Distribution Width, measure of variability in MCV, used to differentiate causes of anemia	44,281	29	21	72
%SAT	Iron	Transferrin saturation, measure of available iron transport capacity	10,180	4	3	75
SedRat	Inflammatory markers	Erythrocyte Sedimentation Rate (ESR), non-specific marker of inflammation	13,945	5	5	100
TIBC	Iron	Total Iron Binding Capacity, measure of iron transport capacity, used to	10,397	1	1	100

		calculate transferrin saturation				
TProt	Liver function	Total Protein in blood	38,352	2	2	100
Trigs	Lipid panel	Triglycerides, tested as part of cholesterol panels	23,963	73	63	86
Troponin	Cardiac markers	Troponin I, heart protein released by damage	10,106	0	0	N/A
TSH	Thyroid function	Thyroid Stimulating Hormone, test of thyroid function and feedback	27,441	1	1	100
UCrea	Renal function	Urine creatinine, measure of kidney function	10,522	0	0	N/A
UricA	Gout	Uric acid, nucleotide breakdown product elevated in gout	7,429	17	14	82
Vi-B12	Nutrition	Vitamin B12, used in DNA synthesis	12,506	7	7	100
Vit-D	Nutrition	Vitamin D storage form, regulates calcium and phosphorus	12,250	6	6	100
WBC	Complete blood count	White Blood Cell count	46,100	33	27	82
TOTAL				**1313**	**982**	**74.8**

Table 2.2: Summary of Novel findings

					MGI-BioVU Meta-Analysis			BioVU Replication Cohort			
Lab	**SNP**	**Chr:Pos**	**Allele 1**	**Allele 2**	**N**	**Beta**	**P-Value**	**N**	**Beta**	**P-Value**	**Replicated**
AlkP	rs3843738	17:43739194	A	G	39,809	0.04	2.51E-08	22,920	0.01	3.58E-01	No
AlkP	rs73004933	19:19675696	T	C	39,809	0.08	4.47E-09	22,730	0.05	7.14E-03	Yes
ALT	rs112574791	8:145730221	A	G	40,116	0.18	3.02E-08	23,007	0.15	5.80E-04	Yes
Amyl	rs1930212	1:104324819	A	G	10,368	-0.25	1.48E-45	3,573	-0.18	4.69E-09	Yes
Amyl	rs8051363	16:75255217	A	G	10,368	0.10	1.07E-10	3,564	0.09	4.51E-04	Yes
BasoRE	rs386785158	15:70744437	T	C	29,653	0.06	7.94E-13	16,191	0.04	2.10E-04	Yes
Bili	rs855791	22:37462936	A	G	39,890	0.04	2.34E-08	22,918	0.04	1.00E-05	Yes
BUN	rs10516957	4:95949206	T	C	45,922	-0.06	1.35E-08	25,245	0.01	6.11E-01	No
Ca	rs6727384	2:97400324	A	G	46,100	-0.04	5.13E-10	25,200	-0.05	2.06E-07	Yes
Ca	rs2839899	9:80350999	A	G	46,100	0.04	6.76E-09	25,194	0.03	9.47E-03	Yes
Cl	rs1030025	2:103105611	A	T	45,920	0.05	4.68E-10	25,204	0.02	9.16E-02	No
FT4	rs10122824	9:139109861	T	G	15,868	0.07	1.00E-09	9,721	0.07	7.28E-07	Yes
Glucose	rs7607980	2:165551201	T	C	46,027	-0.05	4.27E-09	25,312	-0.04	2.09E-03	Yes
Glucose	rs896854	8:95960511	T	C	46,027	-0.04	1.55E-09	25,311	0.01	3.64E-01	No
Glucose	rs9273364	6:32626302	T	G	46,027	0.05	2.63E-11	24,801	0.05	3.10E-06	Yes
HgbA1C	rs3130628	6:31609272	T	C	17,407	-0.08	1.23E-08	7,340	0.03	3.79E-02	No
HCO3 (CO2)	rs1799913	11:18047255	T	G	45,932	-0.04	5.89E-09	25,219	-0.04	7.82E-07	Yes

HCO3 (CO2)	rs77375846	2:103155075	T	C	45,932	-0.10	9.33E-25	25,217	-0.06	2.78E-05	Yes
IGranRE	rs13284665	9:131513370	A	G	30,683	0.22	6.61E-74	QC Fail	N/A	N/A	No
IGranAB	rs13284665	9:131513370	A	G	30,744	0.13	6.76E-35	QC Fail	N/A	N/A	No
K	rs10039139	5:137164863	T	G	45,941	0.07	8.32E-16	25,211	0.06	1.83E-06	Yes
Lipase	rs9377343	6:96512220	A	G	12,649	-0.10	4.79E-14	5,564	-0.08	3.60E-05	Yes
Lipase	rs8051363	16:75255217	A	G	12,649	0.13	2.00E-20	5,549	0.07	8.39E-04	Yes
MCHC	rs12352830	9:80041132	C	G	46,157	-0.04	4.37E-08	26,243	-0.04	5.77E-05	Yes
MonoRE	rs117358683	12:44145965	A	G	32,594	-0.23	2.69E-08	16,185	0.04	4.07E-01	No
MPV	rs11212635	11:108310702	A	T	40,058	0.04	9.55E-09	17,333	-0.01	3.68E-01	No
TProt	rs8022180	14:103263020	A	G	38,352	0.04	7.24E-10	19,665	0.03	2.63E-03	Yes
Trigs	rs6847598	4:76750356	T	C	23,963	-0.05	1.58E-08	12,526	-0.03	1.48E-02	Yes
TSH	rs12590163	14:105223525	T	C	27,441	-0.05	4.68E-08	17,042	-0.04	6.76E-04	Yes
TSH	rs310766	3:12233482	A	G	27,441	-0.06	1.66E-08	17,079	-0.05	1.42E-05	Yes
TSH	rs9275141	6:32651117	T	G	27,441	0.05	3.47E-09	17,054	0.04	8.64E-04	Yes

Table 2.3: Classification of catalog SNPs for alternate summary statistics

https://drive.google.com/file/d/1hlaLOY0wLQli_V6zvgGXhSk8bviLgxit/view?usp=sharing

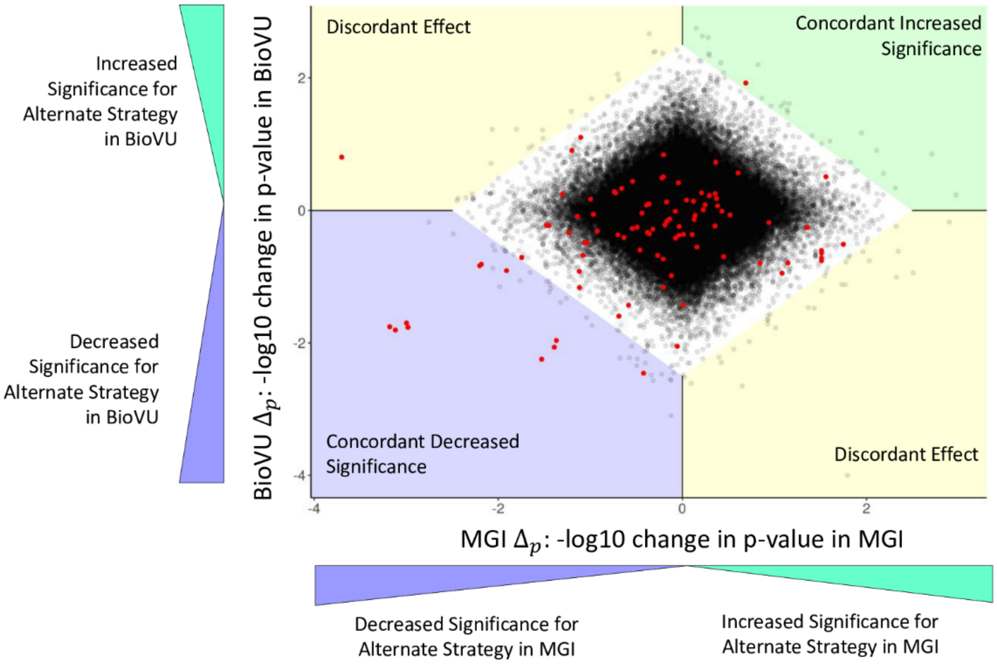
Table 2.4: Classification of catalog SNPs for the comorbidity model, which includes covariates for various lab-altering diseases.

		Comorbidity Model		
Lab	**Testable Catalog SNPs**	**Concordant Increased Significance**	**Concordant Decreased Significance**	**Discordant Effect**
Chol	91	2	5	2
Creat	36	1	3	2
EoAB	31	0	0	0
EoRE	28	0	0	1
HCT	36	2	0	2
HDL	101	15	2	2
Hgb	34	1	0	0
LDL	84	0	7	2
LymphAB	35	2	0	4
LymphRE	20	0	0	0
MCHC	20	2	0	2
MCH	64	1	7	26
MCV	77	9	1	4
MonoAB	43	5	0	1
MPV	84	18	0	5
PLT	102	5	1	4
PMNAB	35	0	2	1
PMNRE	21	0	0	2
RBC	50	2	0	5
RDW	29	0	1	3
Trigs	73	3	3	7
WBC	33	2	2	2

Total	1127	70 (6.2%)	34 (3.0%)	77 (6.8%)

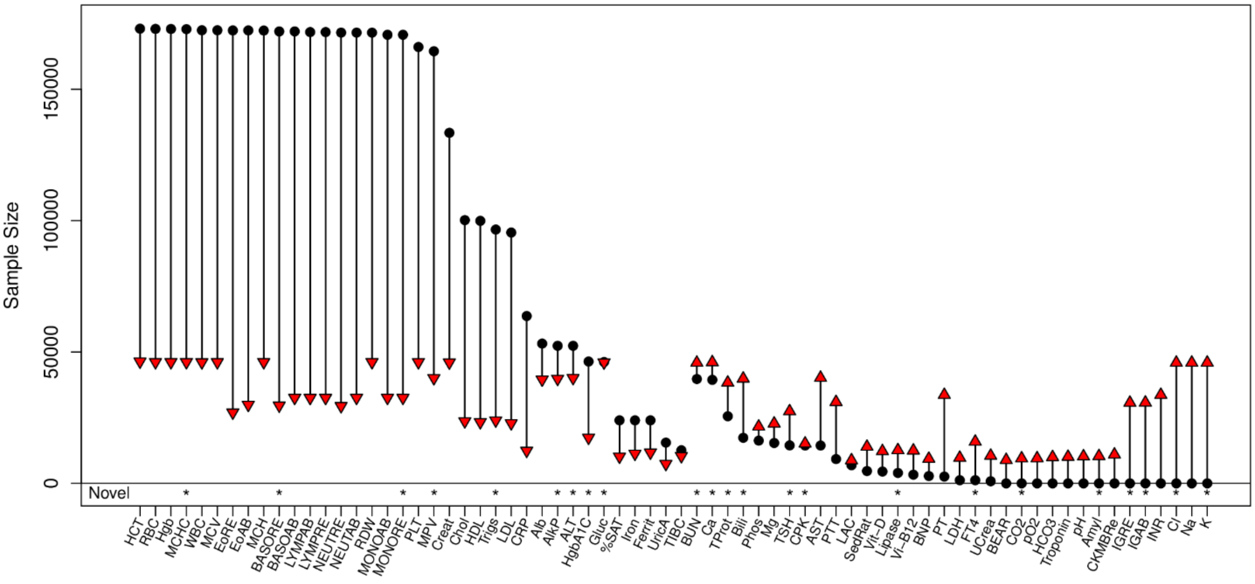
Figures

Figure 2.1: Scatterplot of Δ_p in MGI and BioVU when using the first available measure rather than the mean measurement in a GWAS of Cholesterol level.



Δ_p is the -log fold change in p-value at a SNP for using an alternate analysis, in this case the first available lab measurement. Each dot is a SNP, with red dots indicating GWAS catalog SNPs for the specific lab trait. The white diamond contains 99.9% of SNPs and is used to identify SNPs with the largest changes in p-value due to the alternate analysis. SNPs outside the bounding diamond in the top right (green) quadrant show a concordant increase in significance in both MGI and BioVU, that is, SNPs for which the alternative strategy increases significance in both cohorts. Conversely, SNPs in the bottom left (blue) quadrant show a concordant decrease in significance in both MGI and BioVU. SNPs in either the top left or bottom right (yellow) quadrants have a discordant effect, indicating a large increase in p-value in one cohort but a large decrease in p-value in the second cohort. In this example, one catalog SNP showed a concordant increase in significance when using the first available lab measure, 11 catalog SNPs had a concordant decrease in significance and one SNP had discordant effects. The complete set of scatterplots for each analyzed lab and alternative analysis strategy (summary statistic and comorbidity model) are included in the Supplementary Material. Tables 3 and 4 summarize the movement of catalog SNPs for each lab and analysis strategy.

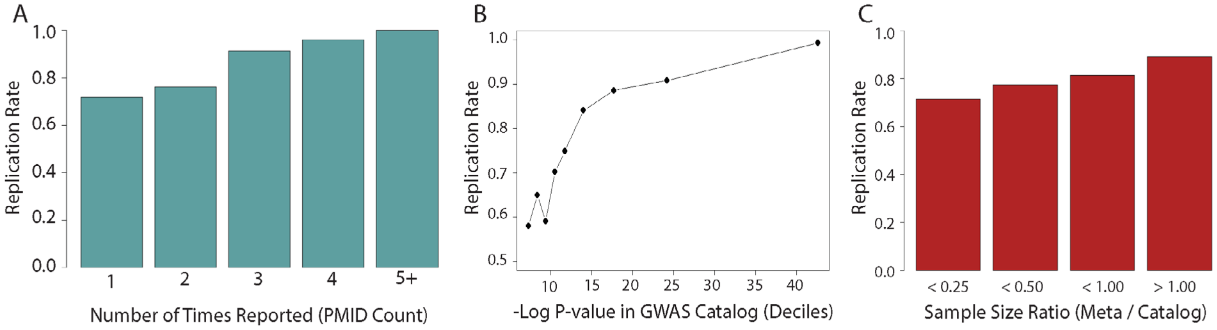
Figure 2.2: Sample sizes for 70 clinical lab traits from the meta-analysis of BioVU and MGI EHRs (red triangles) and the previous largest reported GWAS in a European cohort (black circles).



Our meta-analysis provides the largest GWAS for 34 lab traits, including the first for 14.

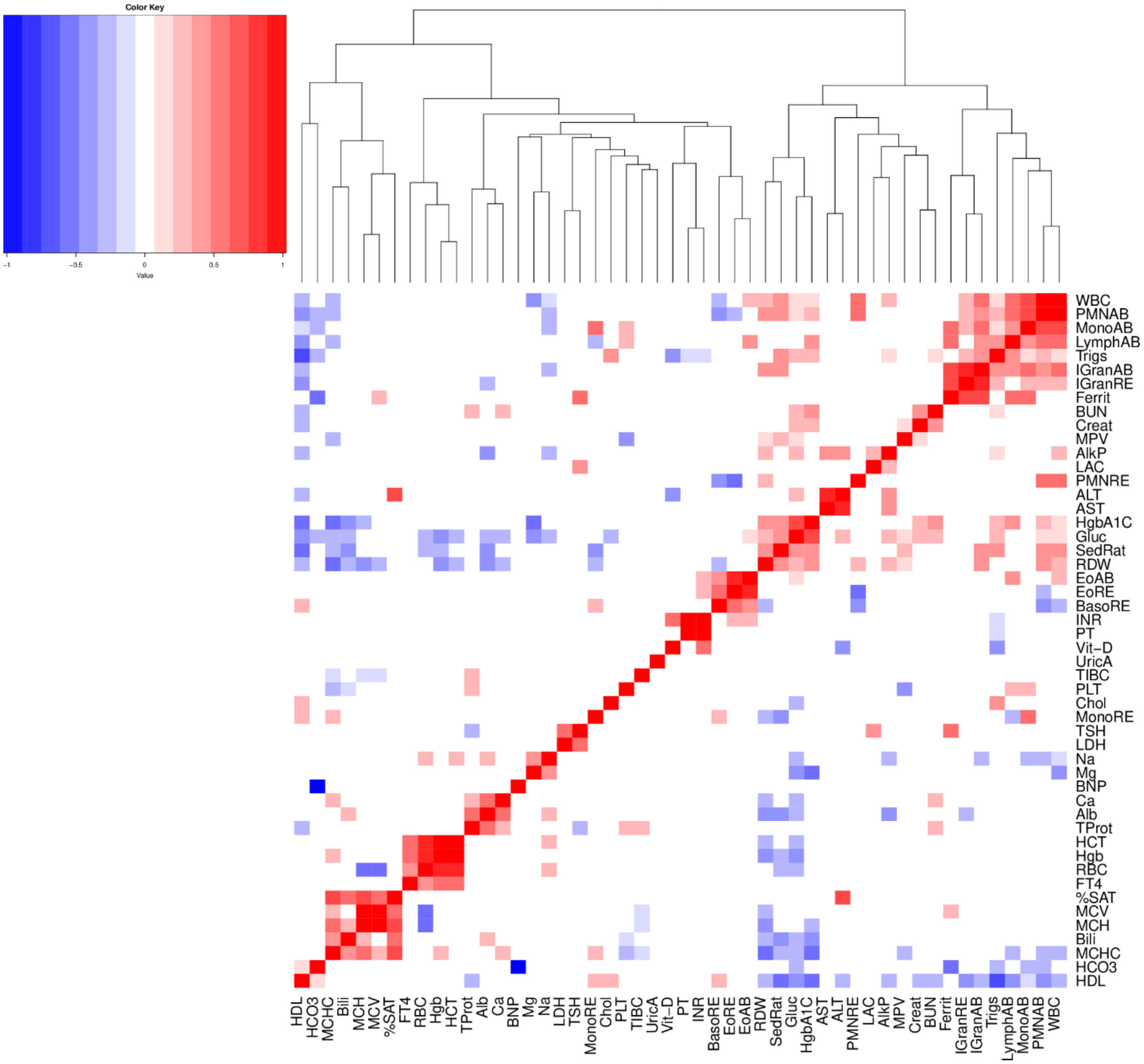
Asterisks along the bottom row indicate labs for which we identified a novel genetic association.

Figure 2.3: Replication rates for GWAS catalog SNPs of clinical labs.



The replication rates increased with (A) the number of times an association was reported in the GWAS catalog, (B) the most significant p-value previously reported for the association, and (C) the ratio of sample size in our meta-analysis to that of the previous largest study.

Figure 2.4: Pairwise genetic correlation of clinical lab traits.



We restricted to labs with heritability of at least 7%. Squares are colored only for correlations having a p-value <0.05 for the null hypothesis of correlation equal to zer

Supporting Information

Supplementary Table 2.5: List of ICD-10 codes used for defining binary trait comorbidities in MGI and BioVU participants for the comorbidity GWAS model.

Disease	Phecode	ICD10
Hypertension	401	I10
Dyslipidemia	272.1	E78
Ischemic heart disease	411	I24, I25
Type 2 diabetes	250.2	E11
Overweight, obesity	278	E66
Tobacco use disorder	318	Z72.0, F17
Osteoarthritis	740	M19
Asthma	495	J45
Epilepsy	345	G40
Hypothyroidism	244	E03
Cerebrovascular disease	433	I67, I69
Heart failure	428	I50
Osteoporosis, osteopenia	743	M80, M81, M83
Chronic airway obstruction	496	J44
Atrial fibrillation	427.2	I48
Arrhythmia NOS	427.5	I49
Chronic kidney disease	585.3	N18

Chronic liver disease and cirrhosis	571	K70, K71, K72, K73, K74, K75, K76, K77
Alcohol use disorder	317	F10
Iron deficiency anemia	280	D50
Type 1 diabetes	250.1	E10
Breast cancer	174	C50
Bipolar	296.1	F31
Rheumatoid arthritis	714	M05
Peripheral vascular disease	443	I73
Prostate cancer	185	C61
Lung cancer	165	C34
Leukemia	204	C91, C92, C93, C94, C95
Non-Hodgkin lymphoma	202	C82, C83, C84, C85, C86, C88, C90, C96
Colorectal cancer	153	C18
Thyroiditis	245	E06
Thyrotoxicosis	242	E05
Aplastic anemia	284	D60, D61
Liver cancer	155	C22
Pancreatic cancer	157	C25
Endometrial cancer	182	C54
Hodgkin disease	201	C81
Cervical carcinoma	180.1	C53

Ovarian carcinoma	184.11	C56
Gastric cancer	151	C16
Esophageal cancer	150	C15
Gallbladder and cholangiocarcinoma	159.3	C23

Supplementary Table 2.6: Table of 1,313 SNPs extracted from the GWAS Catalog based on prior associations with the lab traits and SNPs considered in this study. These associations have been reported at least once in a mixed-sex, adult, European-predominant population not selected for the presence of any disease.

https://drive.google.com/file/d/1LLEBNEHQx8WAhvA-iThxc46xt2dRD-AD/view?usp=sharing

Supplementary Table 2.7: Comparison of GWAS results based on the Area Under the Curve (AUC) summary statistic and the default mean value summary statistic.

Lab	**Normal Range**	**Units**	**Number of Catalog SNPs**	**Smaller p-value for AUC**	**Larger p-value for AUC**	**χ^2 test for p-value change**	**Median fold change**
Chol	(-Inf, 200)	mg/dL	92	20%	80%	5.3E-09	5.5
Creat	(0.7, 1.3)	MG/DL	36	25%	75%	2.7E-03	11.1
EoAB	(-Inf, 0.8)	K/MM3	31	35%	65%	1.1E-01	14.9
EoRE	(-Inf, 6)	%	28	36%	64%	1.3E-01	4.4
HCT	(39, 50.2)	%	36	28%	72%	7.7E-03	12.4
HDL	(40, Inf)	mg/dL	102	13%	87%	5.3E-14	44.1
Hgb	(13.5, 17)	g/dL	34	32%	68%	4.0E-02	2.2
LDL	(-Inf, 100)	mg/dL	85	19%	81%	9.0E-09	8.1
LymphAB	(0.8, 5)	K/MM3	35	26%	74%	4.1E-03	9.3
LymphRE	(20.5, 45.5)	%	20	35%	65%	1.8E-01	4.5
MCH	(27, 32)	pg	64	6%	94%	2.6E-12	5744.8
MCHC	(32, 36)	g/dL	20	10%	90%	3.5E-04	1119.8
MCV	(81, 99)	fl	77	9%	91%	7.0E-13	2916.6
MonoAB	(0.1, 1)	K/MM3	43	12%	88%	4.8E-07	258.4
MPV	(9, 12.2)	fl	84	6%	94%	6.8E-16	13873.8
PLT	(150, 400)	K/MM3	102	10%	90%	4.7E-16	37.4
PMNAB	(1.8, 10.1)	K/MM3	35	29%	71%	1.1E-02	2.1
PMNRE	(43, 65)	%	21	33%	67%	1.3E-01	1.3
RBC	(3.9, 5.3)	M/MM3	50	20%	80%	2.2E-05	5.8
RDW	(11.5, 15)	%	29	14%	86%	9.6E-05	13.2
Trigs	(-Inf, 150)	mg/dL	74	15%	85%	1.5E-09	77.9
WBC	(4, 10)	K/MM3	33	15%	85%	6.2E-05	9.7

Supplementary Text 2.1

GWAS of Area Under the Curve summary statistics of lab traits

We performed a GWAS of lab traits in MGI samples using an Area Under the Curve (AUC) approach to summarize the longitudinal lab measurements. We performed the analysis on the 22 lab traits with 20+ catalog SNPs and compared performance to GWAS with arithmetic mean as the outcome. To construct the AUC statistic, we defined clinically relevant thresholds for each lab trait to differentiate normal variation in measurements from measurements potentially indicative of underlying disease (Table S2, below). The thresholds were based on Normal Range criteria in the EHR and clinical guidance of co-authors. A lab trait could have either an upper threshold, a lower threshold or both depending on the clinical use for diagnosis. For example, the normal range for Red Blood Cell Count (RBC) trait had a lower threshold of 3.9 M/MM3 and an upper threshold of 5.3 M/MM3 since both low and high RBC measurements can be indicative of health problems. In contrast, Low-Density Lipoprotein (LDL) had only an upper threshold of 100 mg/dL since high values of LDL are clinically relevant for disease diagnosis.

We computed individual-level summary statistics that account for both the magnitude and duration of time that longitudinal lab measurements were outside the normal range thresholds. For a given lab trait, let t_l be the lower threshold and t_u be the upper threshold for clinical relevance. Let m_{ij} be the j^{th} lab measurement in the i^{th} subject. Any measurement satisfying $t_l \leq m_{ij} \leq t_u$ is therefore within the "Normal" range of measurements for the given lab trait. Next, define

$$y_{ij} = \begin{cases} m_{ij} - t_u & \text{if } m_{ij} > t_u \\ 0 & \text{if } t_l \leq m_{ij} \leq t_u \\ m_{ij} - t_l & \text{if } m_{ij} < t_l \end{cases} .$$

The quantity y_{ij} is therefore equal to zero if the measurement m_{ij} is in the Normal range and equal to the amount outside the Normal range otherwise. Notably, y_{ij} is positive if the measurement m_{ij} is above the upper threshold and negative if it is below the lower threshold. Thus measurements above the upper threshold accumulate positive area and measurements below the lower threshold accumulate negative area. We set $t_l = -\infty$ for labs with only an upper threshold and $t_u = \infty$ for labs with only a lower threshold.

We computed the accumulated Area Under the Curve statistic for the i^{th} sample based on the Trapezoidal Method as follows:

$$AUC_i \ = \sum_{j=1}^{n_i-1} \frac{1}{2} \Delta_{ij} \ (y_{ij} + y_{ij+1}),$$

where n_i is the number of measurements for the i^{th} sample and Δ_{ij} is the time (in days) between the j^{th} and $(j+1)^{st}$ measurements for the i^{th} sample. For an individual with only a single measurement, we defined the AUC as equivalent to the value ofy_{i1}. We performed GWAS of the AUC values using the same procedure described in the Methods section of the main text.

We found that the AUC-based GWAS performed poorly compared to the standard GWAS of mean trait value based on change in p-values for GWAS catalog SNPs (Table S2). For each lab trait, we computed the proportion of catalog SNPS with smaller p-values (increased significance) and larger p-values (decreased significance) in the AUC GWAS. Assuming the AUC statistic and the mean statistic are equally powerful for summarizing the longitudinal lab measures, we expect that p-values for AUC GWAS will increase for 50% of catalog SNPs on average and decrease for approximately 50% solely due to chance. We found that for all 22 lab traits tested, most catalog SNPs had larger p-values for the analysis based on the AUC summary statistic. That is, the AUC statistic resulted in p-values that with reduced significance. The imbalance was quite extreme: 9 labs traits had >80% of catalog SNPs increase in magnitude and

18 lab traits had p<0.05 for a χ^2 test for equal proportions of catalog SNPs with increasing and decreasing significance. Further, we computed the median fold change across all catalog SNP p-values for each, where fold change greater than 1 indicates a larger p-value for the AUC analysis. All 22 lab traits have median fold change >1 further emphasizing the overall reduced performance of our AUC statistic compared to the standard mean statistic.

The AUC style statistic presents several attractive features for summarizing complex longitudinal lab data; however, we found our implementation to be poor in comparison the basic mean value. We suggest some key limitations of the AUC statistic when applied to the lab measurements. First, restricting the AUC to area outside the clinically thresholds is akin to a censoring natural trait variation and reducing the effective sample size to patients with non-Normal measurements. As an example, 14.5% of samples in the LDL GWAS had AUC values of 0 because none of the LDL measurements for these sample were above the clinical threshold. The use of clinical thresholds is further complicated by the potentially subjective nature of their selection. Second, the AUC statistic can be unduly influenced by individual outlier measurements, particularly those that occur far in time from other measurements. Because the AUC statistic accounts for time between measurements through the Δ_{ij} term, a single measurement outside clinical thresholds is dramatically upweighted if no other measures are taken closely in time afterward. This property is particularly problematic because EHR lab measurements represent a highly imbalanced study design in which successive measurements can routinely occur far apart in time. The AUC statistic might be more effective in a balanced study design with similar numbers of lab measurements at fixed time intervals.

Despite the poor performance here, further refinement in the implementation of an AUC-style summary statistic for EHR lab data could produce more favorable results and is an area for future research.

Supplementary Figure 2.5: The following set of scatterplots show the -log$_{10}$ fold changes in p-value at individual SNPs when comparing GWAS of our default summary statistic (mean) to GWAS based on an alternative statistic (median, maximum or first available).

https://drive.google.com/file/d/1KtCQ2-BKEas6rudOXJfgZZTq9PrzoKkP/view?usp=sharing

Please refer to the Methods section for a complete description. The x-axis corresponds the fold changes for the SNP in MGI and the y-axis corresponds to the fold changes for BioVU. Positive log-fold changes indicate that the alternative statistic yielded a smaller (more significant) p-value than using the mean as a summary statistic. The upper-right (green) quadrant plots SNPs that decreased in p-value in both cohorts for the alternative statistic. The lower-left (blue) quadrant plots SNPs that increased in p-value in both cohorts. The two remaining quadrants indicate SNPs with discordant changes in p-value between the cohorts. GWAS catalog SNPs are plotted in red, novel SNPs for a given lab (if applicable) are plotted in purple, and the remaining SNPs are LD-pruned (for plotting convenience) and plotted in black. The white diamond displays an empirical null distribution of fold changes for non-associated SNPs. The first 22 pages display the three alternative summary statistics (maximum value, median value, and first available measurement) for a single lab. The following six pages contain the analogous plots showing log fold change in p-values for the comorbidity model, which includes binary covariates for various comorbid diseases with the potential to impact lab measures, to a default analysis that does not account for comorbidities.

References

Alexander,D.H. *et al.* (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome research*, **19**, 1655–64.

Astle,W.J. *et al.* (2016) The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell*, **167**, 1415-1429.e19.

Bodenreider,O. (2008) Issues in mapping LOINC laboratory tests to SNOMED CT. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 51–5.

Bulik-Sullivan,B. *et al.* (2015) An atlas of genetic correlations across human diseases and traits. *Nature Genetics*, **47**, 1236–1241.

Bulik-Sullivan,B.K. *et al.* (2015) LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature genetics*, **47**, 291–5.

Buniello,A. *et al.* (2019) The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic acids research*, **47**, D1005–D1012.

Bycroft,C. *et al.* (2017) Genome-wide genetic data on ~500,000 UK Biobank participants. *bioRxiv*, 166298–166298.

Carolina,N. and Carolina,S. (2013) Adoption of Electronic Health Record Systems among U.S. Non-federal Acute Care Hospitals. *ONC Data Brief, no.35.*, 2008–2012.

Chen,S.N. *et al.* (2005) A Common PCSK9Haplotype, Encompassing the E670G Coding Single Nucleotide Polymorphism, Is a Novel Genetic Marker for Plasma Low-Density Lipoprotein Cholesterol Levels and Severity of Coronary Atherosclerosis. *Journal of the American College of Cardiology*, **45**, 1611–1619.

Chertow,B.S. *et al.* (1974) A Biochemical Profile of Abnormalities in Hypothyroidism. *Am J Clin Pathol*, **61**, 785–788.

Das,S. *et al.* (2016) Next-generation genotype imputation service and methods. *Nature genetics*, **48**, 1284–1287.

Devlin,B. *et al.* (2001) Genomic control, a new approach to genetic-based association studies. *Theoretical population biology*, **60**, 155–66.

DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium *et al.* (2014) Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nature genetics*, **46**, 234–44.

FinnGen FinnGen. *FinnGen Documentation of R3 release*.

Fritsche,L.G. *et al.* (2018) Association of Polygenic Risk Scores for Multiple Cancers in a Phenome-wide Study: Results from The Michigan Genomics Initiative. *The American Journal of Human Genetics*, **102**, 1048–1061.

Gagliano Taliun,S.A. *et al.* (2020) Exploring and visualizing large-scale genetic associations by using PheWeb. *Nature Genetics*, **52**, 550–552.

Gao,X. *et al.* (2017) Thyroid hormone receptor beta and NCOA4 regulate terminal erythrocyte differentiation. *Proc. Natl. Acad. Sci. U.S.A.*, **114**, 10107–10112.

Hanauer,D.A. *et al.* (2015) Supporting information retrieval from electronic health records: A report of University of Michigan’s nine-year experience in developing and using the Electronic Medical Record Search Engine (EMERSE). *Journal of biomedical informatics*, **55**, 290–300.

Inchley,C.E. *et al.* (2016) Selective sweep on human amylase genes postdates the split with Neanderthals. *Scientific reports*, **6**, 37198.

Iyengar,S.K. *et al.* (2015) Genome-Wide Association and Trans-ethnic Meta-Analysis for Advanced Diabetic Kidney Disease: Family Investigation of Nephropathy and Diabetes (FIND). *PLoS genetics*, **11**, e1005352.

Kanai,M. *et al.* (2018) Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. *Nature Genetics*, **50**, 390–400.

Kang,H.M. (2014) EPACTS: efficient and parallelizable association container toolbox.

Kichaev,G. *et al.* (2019) Leveraging Polygenic Functional Enrichment to Improve GWAS Power. *American journal of human genetics*, **104**, 65–75.

Klarin,D. *et al.* (2018) Genetics of blood lipids among ~300,000 multi-ethnic participants of the Million Veteran Program. *Nature genetics*, **50**, 1514–1523.

Kooner,J.S. *et al.* (2011) Genome-wide association study in individuals of South Asian ancestry identifies six new type 2 diabetes susceptibility loci. *Nature genetics*, **43**, 984–9.

Krokstad,S. *et al.* (2013) Cohort Profile: the HUNT Study, Norway. *Int J Epidemiol*, **42**, 968–977.

Kullo,I.J. *et al.* (2010) A genome-wide association study of red blood cell traits using the electronic medical record. *PloS one*, **5**.

Larsson,A. *et al.* (2015) The state of point-of-care testing: a european perspective. *Upsala Journal of Medical Sciences*, **120**, 1–10.

McCarty,C.A. *et al.* (2011) The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC medical genomics*, **4**, 13.

McDermott,M.T. (2009) In the clinic. Hypothyroidism. *Ann. Intern. Med.*, **151**, ITC61.

Moon,S. *et al.* (2019) The Korea Biobank Array: Design and Identification of Coding Variants Associated with Blood Biochemical Traits. *Scientific reports*, **9**, 1382.

Morris,A.P. *et al.* (2012) Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature genetics*, **44**, 981–90.

Nagai,A. *et al.* (2017) Overview of the BioBank Japan Project: Study design and profile. *Journal of Epidemiology*, **27**, S2–S8.

Nichols,J.H. (2011) Blood Glucose Testing in the Hospital: Error Sources and Risk Management. *Journal of Diabetes Science and Technology*, **5**, 173–177.

Payne,R.B. *et al.* (1973) Interpretation of serum calcium in patients with abnormal serum proteins. *Br Med J*, **4**, 643–646.

Purcell,S. *et al.* (2007) PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, **81**, 559–575.

Roden,D. *et al.* (2008) Development of a Large-Scale De-Identified DNA Biobank to Enable Personalized Medicine. *Clinical Pharmacology & Therapeutics*, **84**, 362–369.

Rosendahl,J. *et al.* (2018) Genome-wide association study identifies inversion in the CTRB1-CTRB2 locus to modify risk for alcoholic and non-alcoholic chronic pancreatitis. *Gut*, **67**, 1855–1863.

Shioji,K. *et al.* (2004) Genetic variants in PCSK9 affect the cholesterol level in Japanese. *Journal of Human Genetics*, **49**, 109–114.

Shirley,M.D. *et al.* (2013) Sturge-Weber syndrome and port-wine stains caused by somatic mutation in GNAQ. *The New England journal of medicine*, **368**, 1971–9.

Sun,B.B. *et al.* (2018) Genomic atlas of the human plasma proteome. *Nature*, **558**, 73–79.

Tai,M.M. (1994) A Mathematical Model for the Determination of Total Area Under Glucose Tolerance and Other Metabolic Curves. *Diabetes Care*, **17**, 152–154.

Usher,C.L. *et al.* (2015) Structural forms of the human amylase locus and their relationships to SNPs, haplotypes and obesity. *Nature genetics*, **47**, 921–5.

Verma,A. *et al.* (2018) PheWAS and Beyond: The Landscape of Associations with Medical Diagnoses and Clinical Measures across 38,662 Individuals from Geisinger. *The American Journal of Human Genetics*, **102**, 592–608.

Voight,B.F. *et al.* (2010) Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nature genetics*, **42**, 579–89.

Wei,W.-Q. *et al.* (2017) Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record. *PloS one*, **12**, e0175508.

Wei,W.-Q. and Denny,J.C. (2015) Extracting research-quality phenotypes from electronic health records to support precision medicine. *Genome medicine*, **7**, 41.

Willer,C.J. *et al.* (2013) Discovery and refinement of loci associated with lipid levels. *Nature genetics*, **45**, 1274–1283.

Willer,C.J. *et al.* (2010) METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*, **26**, 2190–2191.

Wolever,T.M.S. and Jenkins,D.J.A. (1986) The use of the glycemic index in predicting the blood glucose response to mixed meals. *Am J Clin Nutr*, **43**, 167–172.

Xue,A. *et al.* (2018) Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. *Nature communications*, **9**, 2941.

Chapter 3 Sparse Bayesian Mendelian Randomization Increases Power for Estimation of Causal Effects

Introduction

Genome-wide association studies (GWAS) have identified numerous disease-associated loci in recent years, advancing our understanding of disease etiology (Visscher *et al.*, 2017). GWAS have contributed to discovering novel biological mechanisms in cardiovascular disease (Teslovich *et al.*, 2010; Willer *et al.*, 2013), type 2 diabetes (Mahajan *et al.*, 2018), and psychiatric phenotypes (Ripke *et al.*, 2014), among others. The advent of large-scale biobanks, which link electronic health record (EHR) derived phenotypes to genotype data, has accelerated these analyses, enabling simultaneous association studies across thousands of phenotypes at sample sizes that approach 500,000 (Bycroft *et al.*, 2017; Klarin *et al.*, 2018).

Despite the preponderance of disease-associated loci and steadily increasing sample sizes, our understanding of the causal links between phenotypes remains limited in comparison. This foundational epidemiological work has been limited by the challenges of implementing study designs that permit causal inference, such as randomized experiments. Observational data, which are easier to collect, are far more common, but make causal inferences more challenging. However, under certain circumstances, Mendelian randomization (MR) methods enable causal inference on observational data available in biobank. The premise and epistemology of MR refers to Mendel's law of independent assortment, suggesting that random inheritance of paternal or maternal alleles is akin to a randomized experiment, which would preclude the effect of confounding and reverse causation, which are otherwise ubiquitous concerns when using observational data for causal inference.

MR is an instrumental variable (IV) analysis employing observed germline genetic variation as an instrument for an exposure (Davey Smith and Ebrahim, 2003). An instrument for an exposure is typically constructed using the single-variant association statistics from an association study in order to estimate the causal effect of the exposure on an outcome after employing MR machinery, typically a form of regression analysis. MR requires three assumptions: relevance, independence, and exclusion. The relevance assumption requires that the variants are truly associated with the exposure, which can be assessed through common regression test-statistics. The independence assumption requires that the instrument variants are not associated with any unmeasured confounders, and the exclusion assumption requires that the effect of the instrumental variants on the outcome is completely mediated through the exposure. The independence and exclusion assumptions are challenging to assess, given that studies rarely contain all possible covariates and confounders, meaning that they cannot be explicitly tested. The analysis of complex traits complicates assumption validation in MR, as often confounders that may be associated with genotypes are unknown, including those related to population stratification. Recent MR methods have posited statistical methods that use mixture modeling assumptions to assess whether there is evidence of pleiotropy (Qi and Chatterjee, 2019; Morrison *et al.*, 2020).

Although the first assumption is readily tested by examining the exposure GWAS p-values and effect sizes, variable selection in a high-dimensional setting remains challenging, and single-variant summary statistics rarely incorporate shrinkage information that may yield better selection. In practice, many variants are in linkage disequilibrium (LD), which leads to challenging instrument selection. Many MR methods require independent variants (Zhu *et al.*, 2016; Bowden *et al.*, 2016), necessitating choosing among seemingly equivalent variants for the purposes of IV analysis. Many GWAS-associated loci result in large association peaks (Zhang and Lupski, 2015) in non-coding regions, which rarely conclusively implicate a single variant conclusively. However, modern advances in MR methods enable simultaneous use of correlated instruments (Yuan *et al.*, 2020), leading to increased power. Nevertheless, instrument selection remains a challenging step in MR.

Imputation methods and whole-genome sequencing (WGS) have enabled association mapping at a fine resolution, increasing GWAS in both sample size and variant density. The greater density of variants includes a swath of low-frequency variants and variants in low-LD. Indeed, recent analysis from the GIANT consortium (Wainschtein *et al.*, 2019) suggests that low-LD and low-MAF variants contribute a substantial portion to heritability of complex traits. Phenotypes with substantial heritability contributions from these variants may benefit from assuming a sparse genetic architecture, given that their effects are more poorly tagged by other variants than common variants in high-LD regions. The abundance of these low-frequency and low-LD variants in GWAS summary statistics will continue to increase with modern imputation panels (Taliun *et al.*, 2021) and more common WGS. As GWAS data continue to grow faster in variant density than sample size, methods that assume a sparse causal architecture are likely to be beneficial.

Furthermore, MR was initially only applied in single-sample settings where both exposure and outcome were available on the same set of individuals. Recent methods have been introduced that allow for an exposure and outcome from possibly overlapping sets of samples, which are referred to as two-sample MR (Pierce and Burgess, 2013). These methods have provided additional convenience as often the exposure and outcome of interest come from distinct studies. To permit this convenience, methods have been developed to use existing GWAS summary statistics (Zhu *et al.*, 2016) rather than require individual level data from two separate studies, which often presents challenges with data privacy and data sharing. This has necessitated modified likelihoods that must be constructed without access to the underlying individual level data, and typically use sufficient statistics of the data to construct approximate likelihoods (Zhu and Stephens, 2017).

Building on recent advances in MR methodology, we here introduce a sparse Bayesian two-sample MR method, which we term SPARMR (SPARse Mendelian Randomization), which has improved power and decreased estimation error when the genetic architecture is sparse, compared to methods that assume a dense causal architecture. SPARMR can be applied to GWAS summary statistics from separate studies and performs simultaneous causal effect estimation and testing. Unlike most applications of sparse regression to genetic association studies, SPARMR does not use a spike-and-slab prior, which has a combinatorial computational complexity, but instead uses a horseshoe prior from the family of continuous shrinkage priors. The horseshoe prior has previously been applied in MR with application to a pleiotropy term (Berzuini *et al.*, 2020), but to our knowledge has not been applied for instrument selection. Unlike the spike-and-slab prior, continuous-scale shrinkage priors facilitate the use of efficient Hamiltonian Monte-Carlo samplers. Continuous-scale shrinkage priors have also been applied in the context of polygenic risk scores (Ge *et al.*, 2019). Here we offer a convenient application interface to the model, which is implemented with Tensorflow Probability (Dillon *et al.*, 2017). We demonstrate the effectiveness of SPARMR in a two-sample setting through simulations derived from 1000 Genomes Project (Auton *et al.*, 2015). We apply our method to investigate the effects of lipids on cardiovascular outcomes in UK Biobank, and conclude with a discussion.

Materials and Methods

SPARMR Overview

We denote the exposure as *X*, and the outcome as *Y*. We are interested in the causal effect of *X* on *Y*, using a set of bi-allelic genotypes *G* (n x p) to construct an instrument. Let L_X, L_y be the matrices of non-genotype covariates that may include principal components or other relevant covariates for *X* and *Y*. We model the joint distribution of *X* and *Y* as

$$f\left(\left(\begin{smallmatrix}\tilde{Y}=y\\ \tilde{X}=x\end{smallmatrix}\right)\middle|G\right)=N_2\left(G\begin{pmatrix}\theta\alpha\\ \theta\end{pmatrix},\begin{pmatrix}\sigma_y^2 & \rho\\ \rho & \sigma_x^2\end{pmatrix}\right)$$

Where $\tilde{Y}=Y-L_y\left({L_y}^tL_y\right)^{-1}L_y^tY$, and. $\tilde{X}=X-L_{\mathrm{x}}\left({L_x}^tL_x\right)^{-1}L_x^tX$. We interpret α as the effect of *X* on *Y* and θ as the effect of G on $\tilde{X}$. Let σ_x^2 be the residual variance of $\tilde{X}$ after conditioning on G, and σ_y^2 the residual variance of $\tilde{Y}$. Let ρ be the residual covariance between *X* and *Y* after conditioning on G.

In practice, we do not observe the individual sample values from *G, X,* or, *Y*, but rather sufficient statistics of the data, and we instead formulate an alternative summary statistic likelihood. We use summary statistics (coefficient point estimates and standard errors) from the two GWAS of *X* and *Y* and denote the summary statistics as $\left(\widehat{\beta_X},\widehat{s_x}\right)$ and $\left(\widehat{\beta_y},\widehat{s_y}\right)$. We assume the coefficients come from the marginal association between the matrix of possibly correlated genotypes, *G*, and the phenotype (*X* or *Y*). We assume that the marginal association statistics have already been adjusted for population stratification and cryptic relatedness through the inclusion of principal components (Price *et al.*, 2006; Novembre and Stephens, 2008) or application of a mixed model (Kang *et al.*, 2010; Loh, 2015; Zhou *et al.*, 2018). We assume that other relevant covariates have already been included in the respective GWAS.

If the summary statistics come from two studies with no sample overlap, then it is reasonable to assume $\rho = 0$. For the purposes of exposition, we assume no sample overlap. In this setting, the bivariate Gaussian likelihood can be expressed as $L\left(\tilde{Y},\, \tilde{X}\middle|G\right) = exp\left(-\frac{1}{2}\left[\frac{(\tilde{Y}-G\alpha\theta\,)^2}{\sigma_y^2}+\frac{(\tilde{X}-G\theta\,)^2}{\sigma_x^2}\right]\right)(2\pi)^{-1}\left(\sigma_y^2\sigma_x^2\right)^{-1/2}$. Expanding out the exponentiated term yields $-\frac{\widetilde{y^t}\tilde{y}-2\widetilde{y^t}G\alpha\theta+\alpha^2\theta^tG^tG\theta}{2\sigma_y^2}-\frac{\widetilde{x^t}\tilde{x}-2\widetilde{x^t}G\theta+\theta^tG^tG\theta}{2\sigma_x^2}$. Following (Yang *et al.*, 2012), we can approximate these terms using summary statistics derived from the GWAS results and external genome population databases. We assume the genotypes have been centered at their mean value, $2\, *\, AF$, where *AF* is the allele frequency of the alt-allele. Let $G^tG = n\widehat{Cov(G)} = n\Delta^{1/2}R\Delta^{1/2}$, where *R* is the variant correlation or LD matrix, and the Δ is a diagonal matrix of SNP variances, both of which can be estimated from an external LD reference and minor allele frequency (MAF) database, such as 1000 Genomes (Auton *et al.*, 2015) if the sample population is included in 1000 Genomes. Let $D = diag(G^tG)$ and $\widetilde{y^t}G = D\widehat{B_y}$. We do not estimate the σ_y^2or σ_x^2, but rather approximate them with the expected $\widehat{\sigma^2}$ from the single SNP regression in the exposure and outcome respectively with the smallest p-value. We approximate other terms following (Yang *et al.*, 2012). We use this summary statistic likelihood along with our prior specification to perform MCMC inference to estimate the posterior.

Prior specifications

We assume a horseshoe prior over the genotype effects θ. The classical horseshoe estimator (Carvalho *et al.*, 2010), introduced in the normal means setting, is represented as:

$$\theta_i \sim N(0, \lambda_i^2\tau^2), \lambda_i \sim C^+(0,1), \tau \sim C^+(0,1)$$

Where C^+denotes the Half-Cauchy distribution over R^+. The horseshoe prior includes local scale parameters $\lambda_i \in R^+$, which control the extent to which individual SNPs are shrunk, and a global scale parameter $\tau \in R^+$, which induces shrinkage across all SNPs. The etymology of the name "horseshoe" refers to the observation that the shrinkage induced by the prior looks like a horseshoe – that is, coefficients are either completely shrunk toward 0, or have very little shrinkage. The horseshoe results in a similar shrinkage effect to a spike and slab prior (Supplementary Figure 1).

We motivate the use of a sparse prior by the empirical observation that associated variants are relatively infrequent, even for complex phenotypes. Height, widely considered as the canonical polygenic trait, has 3,290 independent SNPs according to a 2018 meta-analysis of ~700,000 individuals from the GIANT consortium (Yengo *et al.*, 2018). The number of segregating polymorphisms humans very likely exceeds one billion, according to whole-genome analysis from the NHLBI Trans-Omics for Precision Medicine (TOPMed) consortium (Taliun *et al.*, 2021), implying a density of $3.29x10^{-6}$ for the most polygenic trait in terms of independent associated variants, assuming that we already have discovered a comprehensive set of variation associated with height.

Modern theoretical population genetics models also tend to implicate a small number of causal loci relative to the entire genome. The omni-genic model (Boyle *et al.*, 2017), suggests that nearly any gene that is expressed in a trait-relevant tissue is likely to have some causal contribution to a given phenotype. Despite ostensibly implicating a far greater density of causal variants than alternative models of genetic architecture in complex traits, many traits will be sparse in the number of trait-relevant expressed genes compared to the entire genome. This is contrast with the infinitesimal model introduced by RA Fisher (Fisher, 1919), which assumes that the variant effects are exchangeable with respect to a gaussian prior, i.e $\theta_i \sim N(0, \sigma^2)$, implying that causal effects are dense.

Another useful property of the horseshoe distribution is its limited shrinkage of large coefficients compared to lasso regression (Carvalho *et al.*, 2009). Rare variants that abrogate function of disease relevant genes may have large effect in association studies, especially in traits under negative selection (Zuk *et al.*, 2014; Pritchard, 2001). Current priors that assume an infinitesimal model are likely to shrink these large effects more strongly than is necessary (Zhou *et al.*, 2013). As these large effect rare variants are potentially of keen interest towards informing potential therapeutic options, properly incorporating their effects in a Mendelian Randomization setting is useful. Although GWAS of complex traits to date generally have identified few rare variants of large effect (Fuchsberger *et al.*, 2016), this is more likely to occur in the future given the advent of sequencing of large scale biobanks, such as the effort to sequence all UK Biobank participants and the All of Us initiative from the NIH.

On the causal effect parameter α, we place the following mixture prior:

$$\alpha \sim .5 * N\left(0, \sigma_{\alpha_1}^2\right) + .5 * N\left(0, \sigma_{\alpha_2.}^2\right)$$

With the second gaussian representing a “spike” of a null causal effect, akin to a continuous approximation of a spike-and-slab prior. To estimate an inferential statistic, we post-process the MCMC samples to determine the proportion of samples where α was more likely to be generated by the slab $N\left(0, \sigma_{\alpha_1}^2\right)$ distribution. In practice, we set $\sigma_{\alpha_1}^2 = 1.0$ and $\sigma_{\alpha_2}^2 = 0.001$. We used this approximation instead of a traditional discrete latent variable as this approach is precluded by Hamiltonian Monte-Carlo approaches that require a differentiable likelihood. We refer to this statistic as a posterior inclusion probability (PIP), a common inferential statistic in Bayesian methods (Wen *et al.*, 2016; Benner *et al.*, 2016).

The variable selection performed here with a continuous shrinkage prior is facilitated with a Bayesian approach. Several frequentist approaches exist for variable selection, including the lasso, best subsets-regression, and step-wise regression. However, these approaches are limited either by computational complexity (best-subsets) or by challenging statistical hypothesis testing (lasso). As MR is both an estimation and testing problem, inference is among its goals. A Bayesian approach enables both sparse estimation via an estimate of the expectation of the posterior distribution when a sparse prior is used, and inference, through Bayesian model selection.

Prior specification and parameter initialization

In practice, the heavy tails of the C^+ distribution make reliable MCMC sampling challenging. We do not use the C^+distribution, but instead a folded T-distribution with 4 degrees of freedom, which offers similar benefits to the $C^+$and is more amenable to computation. We also fix the global shrinkage parameter τ using heuristics from (Piironen and Vehtari, 2017) based on the expected number of causal variants. We initialize θ_i with a random draw from the horseshoe distribution, and we initialize α using a draw from estimate of the slope of the linear regression of $\widehat{\beta_y}$ on $\widehat{\beta_x}$.

Sampling of Posterior

We sample from the posterior using the No-U-Turn sampler (NUTS) (Hoffman and Gelman, 2014) as implemented in the Tensorflow Probability probabilistic programming language (Dillon *et al.*, 2017), using four separate MCMC chains with 800 burn-in samples each. To asses convergence, we calculate effective sample size (ess) and the $\hat{R}$ statistic for each parameter using the Arviz (Kumar *et al.*, 2019) package.

Posterior SNP summary statistics

Although SPARMR estimates a full posterior distribution for each θ_i, in practice point estimates are frequently of interest. We used the posterior mean of each θ_i as the point estimate. Additionally, similar to the spike-and-slab, we can derive quantities that indicate the extent of shrinkage for each SNP *i.* We can calculate a pseudo inclusion probability by comparing the local shrinkage posterior for a given SNP to a threshold, λ_c, i.e $pseudo - PIP\ =\ P(\lambda_i\ >\ \lambda_c)$. Greater λ_i values indicate less shrinkage.

Simulating summary statistics

We performed simulation to compare SPARMR to other two-sample Mendelian Randomization methods. We used 2,504 individuals from 1000 Genomes Phase 3, from a 0.5 Mb chunk from chr22 in a gene dense region (chr22:37200001-37700000). We split the 2,504 samples into two mutually exclusive sets of 1,252 samples two emulate a two-sample setting. We simulated the phenotypes with multiple genetic architectures, and varied the sparsity, the proportion of genetic variants with non-zero coefficients π, between 0.1% and 1%. That is, we simulate the causal genetic variant effect sizes according to the mixture distribution, $\theta_i \sim (1 - \pi)\delta_0 + \pi N(0, {\sigma_\theta}^2)$.

We fixed the ${h_{exposure}}^2 = \frac{Var(G\theta)}{Var(trait)}$ to 20% for the exposure, a similar heritability to those of many laboratory traits (Sinnott-Armstrong *et al.*, 2021), including lipids and blood cell traits. We varied the causal effect size α between .05 and 1.0 for the type 2 simulations and fixed the causal effect parameter to 0 for the type 1 simulation. Because the variance of the outcome was fixed to 1, this implies that the variance explained by the exposure on the outcome was $\alpha^2 {h_{exposure}}^2$ For each setting, we ran 1,000 simulations, and calculating single-SNP marginal summary statistics for each simulation. In total, we ran 16,000 simulations, with 2,000 of them being type 1 error simulations, setting the maximum false discovery rate for a method is 12.5% (if it rejected the null on every simulation). We did not include a horizontal pleiotropy effect. Many of the SNPs are in high LD with one another, we did not perform LD pruning, and proceeded with correlated instruments.

Estimation of LD

We estimated LD directly on the 1000 Genomes genotypes using all available samples (n=2,504). We calculate the LD matrix using the corrcoef function from the Python numpy package (Harris *et al.*, 2020). As the number of variants exceeded the number of samples, the observed correlation matrix was not positive definite, necessitating further processing. We used the nearPD function from the Matrix package (Bates *et al.*, 2019) to find the 'nearest' positive definite correlation matrix.

Alternative Two-Sample Mendelian Randomization Methods

We compared our method to five other MR methods observed in the literature. We included (1) inverse-variance weighted average (IVW, (Burgess *et al.*, 2013)), (2) Egger regression (Bowden *et al.*, 2015), (3) MRMix (Qi and Chatterjee, 2019), (4) weighted median (Bowden *et al.*, 2016), and (5) Summary Mendelian Randomization (SMR) (Zhu *et al.*, 2016). For IVW, Egger Regression, and weighted median estimator, we used the implementations in the MendelianRandomization R package (Yavorska and Burgess, 2017). Although historically these estimators were constructed with independent SNPs in mind, the MendelianRandomization package provides an option for the user to specify a correlation matrix if correlated SNPs are used, which we have done here. We used the single SNP with the smallest pvalue in the exposure GWAS as the instrument for SMR. Except for SMR, we used the same input SNPs for every method including SPARMR. We determined the input SNPs by including those with a pvalue <

5e-8 in the exposure association summary statistics, as pre-filtering of SNPs based on pvalue is a common practice. The number of input SNPs average 165.2, the median 121, the minimum 10, and the maximum 749.

Briefly, we describe the alternative two-sample methods. IVW performs a weighted linear regression of the outcome coefficients on the exposure coefficients, where the weights are the inverse square standard errors of the outcome coefficients, an approach very similar to two-stage least squares. Egger regression is similar to IVW with the exception that the absolute value of the exposure coefficients is used, and the sign of the outcome coefficients is altered to match the sign of exposure coefficients prior to absolute value transformation. The intercept can be interpreted as an estimator of the average horizontal pleiotropic effect across the included instruments.

As IVW, Egger regression, MRMix, and SMR are frequentist estimators, they produce p-values in contrast to SPARMR, which produces PIPs. To calibrate our comparison, we evaluate all estimators with respect to power and false discovery rate at different thresholds of their respective summary tail measures. This ensures that power evaluations are interpreted within context of matched false positive and false discovery rates.

IVW, Egger, weighted median, and MRMix all either assume no distribution on the causal effects or assume the effects are dense. In contrast, SMR assumes the causal architecture is sparse, which aligns it more closely with SPARMR in this setting. All methods used are capable of application to a two-sample summary statistic setting.

Analysis of the effect Laboratory Trait Values on Coronary Artery Disease

We use summary statistics from a GWAS of laboratory trait values from the Michigan Genomics Initiative (MGI), which have been previously described (Goldstein *et al.*, 2020).

Briefly, MGI is a biobank with samples recruited from the University of Michigan Health System and has samples which are primarily of European ancestry. We included the low-density lipoprotein (LDL) from 11,016 individuals and estimated its effect on coronary artery disease (Phecode 411). For some individuals, multiple LDL measurements were recorded, and in such cases we used the arithmetic mean. We computed GWAS summary statistics using EPACTS (Kang, 2014) including age, sex, and four genetic ancestry PCs as covariates, and inverse normal transformed the phenotype. The genotypes were imputed with the Haplotype Reference Consortium (HRC) panel (Das *et al.*, 2016) and we included all genotypes with imputation Rsq of at least 0.30. We used summary statistics from UK Biobank for coronary artery disease. The UK Biobank has been described previously (Bycroft *et al.*, 2017). These genotypes were imputed with the HRC and the summary statistics were computed with SAIGE (Zhou *et al.*, 2018) to account for case-control imbalance.

Results

We introduce a two-sample sparse Bayesian method for Mendelian Randomization on summary statistics, as described in materials and methods. Our method models multiple correlated instruments and uses a continuous shrinkage prior on the coefficients. We refer to our method as Sparse Mendelian Randomization (SPARMR) and is implemented in a publicly available python package. We investigated the false positive and false negative error frequencies and the empirical false discovery rate of our method. We evaluated the false positive rate (Supplementary Figure 2) and false discovery rate (Supplementary Figure 3) of the methods by multiple decision thresholds stratified by the sparsity of the true causal variants (i.e., the proportion of variants that have non-zero effects on the exposure). The false positive rate of the frequentist methods varied, and SMR was the closest to correct nominal coverage (i.e., if the null was rejected at a p-value of < .05, we expect 5% of the null simulations to reject the null), followed by MRMix. Egger regression, IVW, and weighted median displayed marked type one error inflation, with false positive rates exceeding 25% at a p-value threshold of .001. The false positive rate of SPARMR exceeded 25% at PIP thresholds below 50%. Aligning the decision thresholds of SPARMR with the frequentist methods in terms of false positive rate or false discovery rate enables calibrated comparison of the power of these methods under various settings.

SPARMR is well-powered under sparse settings

We observed the power of these methods under varying causal effect sparsity and effect size. SPARMR, SMR, and weighted median performed similarly in both sparsity settings (Figure 1), with the other methods indicated less power. SMR had the greatest power in the sparsest setting, and its performance decreased as the causal variant density increased. In contrast, the weighted median estimator was more highly powered under the denser setting.

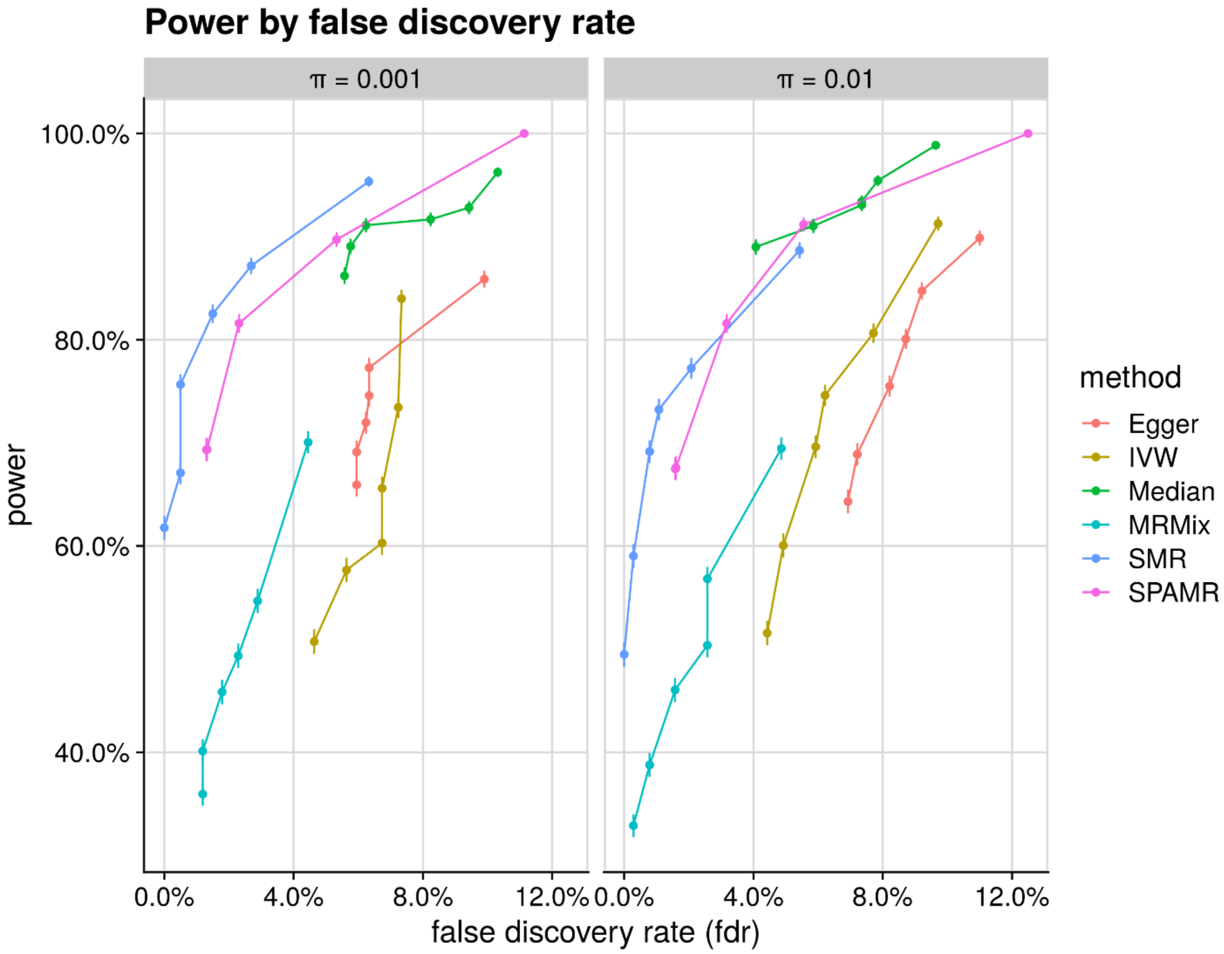


Figure 3.1: Power across different False discovery Rate Thresholds

Power is observed marginalized over multiple causal effect sizes, stratified by causal effect sparsity. 95% CI intervals are displayed with lines at each of the points, the interval calculated with the asymptotic gaussian approximation to the binomial.

Estimation of bias in causal effect estimates

We compared the α estimates from the different MR methods. The frequentist methods report a point estimate for α, while SPARMR produces a posterior distribution. We used the posterior mean as the point estimate for SPARMR. All estimators displayed little bias for α values less than 0.35 (Figure 2). As the α increased, so did bias in the respective estimates. The frequentist estimators were frequently conservative, with SMR displaying the least bias. SPARMR was anti-conservative among the higher α settings, which may reflect biases induced by approximations in the likelihood calculation. However, we briefly remark that in real applications α will often be smaller than these large settings.

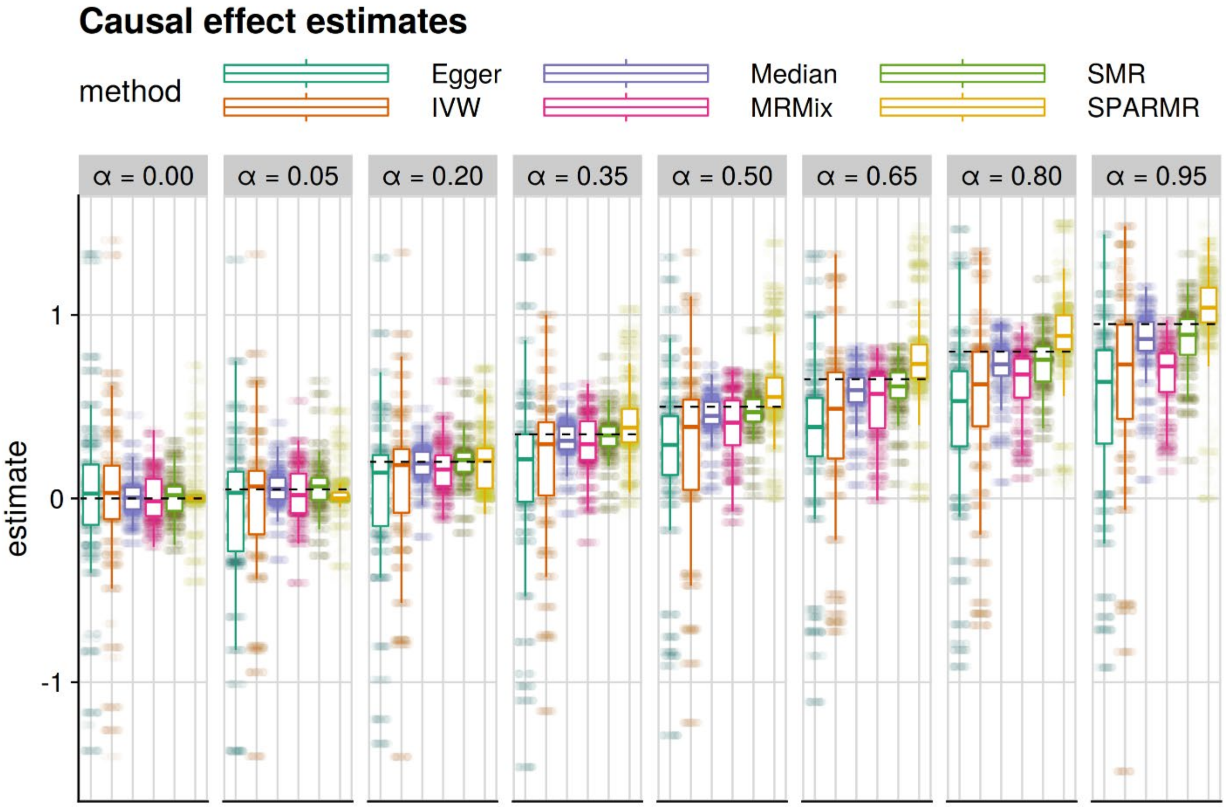


Figure 3.2: Distribution of estimated causal effects stratified by causal effect

Each point is a point estimate from a single simulation, and the box plots display a summary of the distribution within each of the causal effect settings. The results are marginalized across the two sparsity settings. The true causal effect estimate from the simulation is labeled at the top panel and represented by a dashed black line.

Estimating the causal effect of LDLR variation on ischemic heart disease

The causal effect of variation in LDL on variation in ischemic heart disease has been well described (Kathiresan *et al.*, 2008; Willer *et al.*, 2013). We asked whether SPARMR could recapitulate known causal associations of large effect to demonstrate as a proxy for a positive control. However, despite the overall association between LDL and ischemic heart disease, the mechanisms through which genetic variation alters expected LDL levels in adulthood varies, with some loci more well characterized than others. We focus on SNPs within LDLR, a gene which encodes the low-density-lipoprotein receptor and is well described as a locus for LDL variation.

We performed a GWAS of LDL in MGI, controlling for age, sex, and genetic ancestry principal components (see Methods). Variants LDLR were associated with LDL in the MGI cohort at genome wide significance (Supplementary Figure S4), fulfilling the instrument strength assumption of Mendelian Randomization. The T allele of the lead variant rs6511720, which is a ClinVar (Landrum *et al.*, 2016) variant for familial hypercholesterolemia, was associated with a 0.15 decrease in standard deviations of LDL (pvalue = 1.2e-13). For these same variants, we calculated LD directly on the MGI genotypes, and included variants where the MAC exceeded 40. MGI and UK Biobank both primarily have samples of European ancestry, suggesting that MGI genotypes provide an appropriate reference for estimating LD. The estimated correlation matrix was not positive definite, so we used the nearPD function from the R package Matrix (Bates *et al.*, 2019) to find the 'nearest' positive definite correlation matrix. We used pre-computed GWAS summary statistics from UK Biobank that were imputed into the same panel as the MGI data (see Methods). SPARMR estimated a large effect ($\hat{\alpha}$ = 0.78, 95% credible interval: (0.48, 1.09), $R \widehat{=\ 1}.0$) of LDL on ischemic heart disease, consistent with prior reports. The effect sizes can be interpreted as the expected effect of an increase in the exposure by one standard deviation on the log-odds of the outcome.

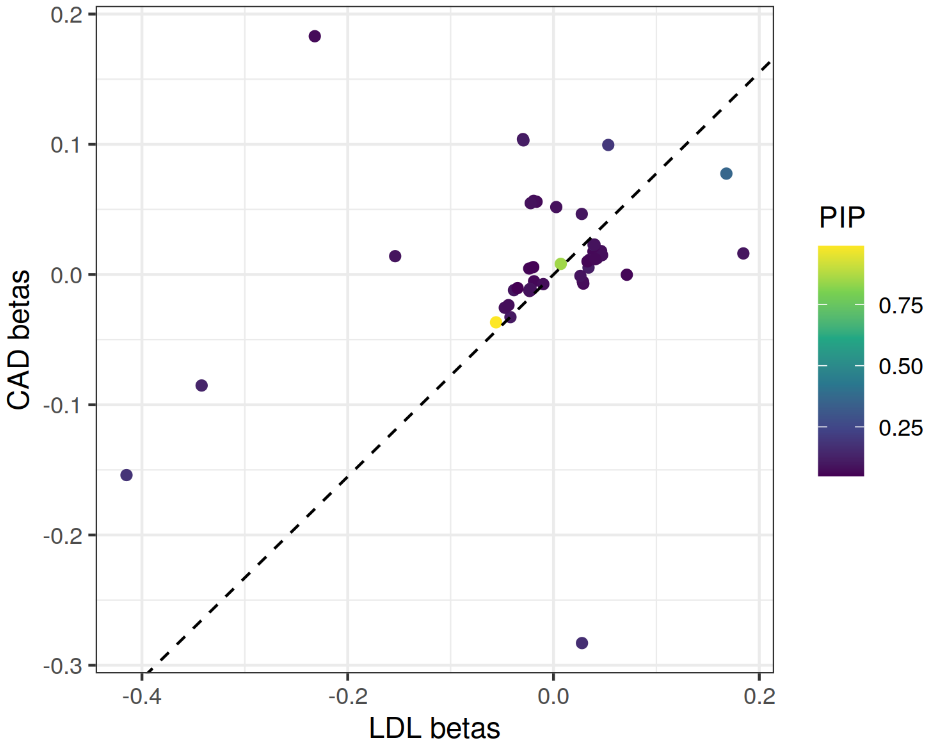


Figure 3.3: SPARMR recapitulates positive control of causal effect of low-density lipoprotein (LDL) on coronary artery disease (CAD).

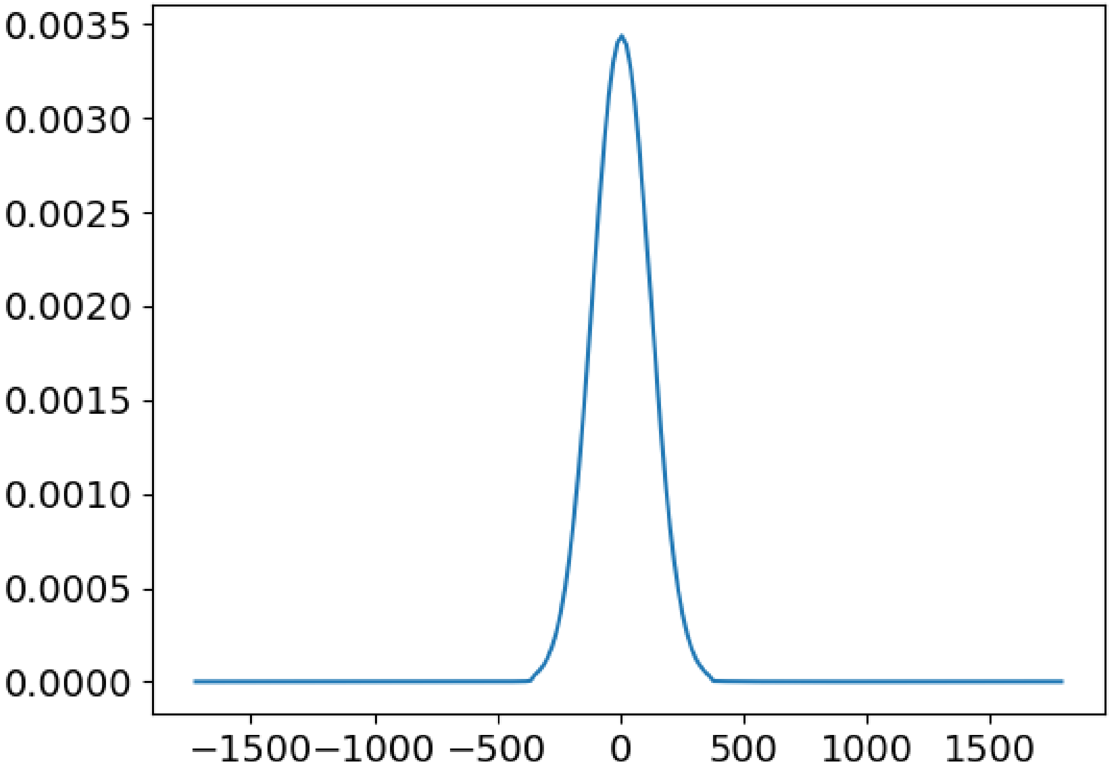
GWAS summary stats for LDL were estimated using the Michigan Genomics Initiative, and summary statistics for CAD are estimated from UK Biobank. The estimate here is applied to 53 common SNPs in LDLR. Dotted line indicates estimated causal effect from SPARMR. Points are colored based on pseudo-PIP estimates from SPARMR, which indicates which SNPs were selected.

Discussion

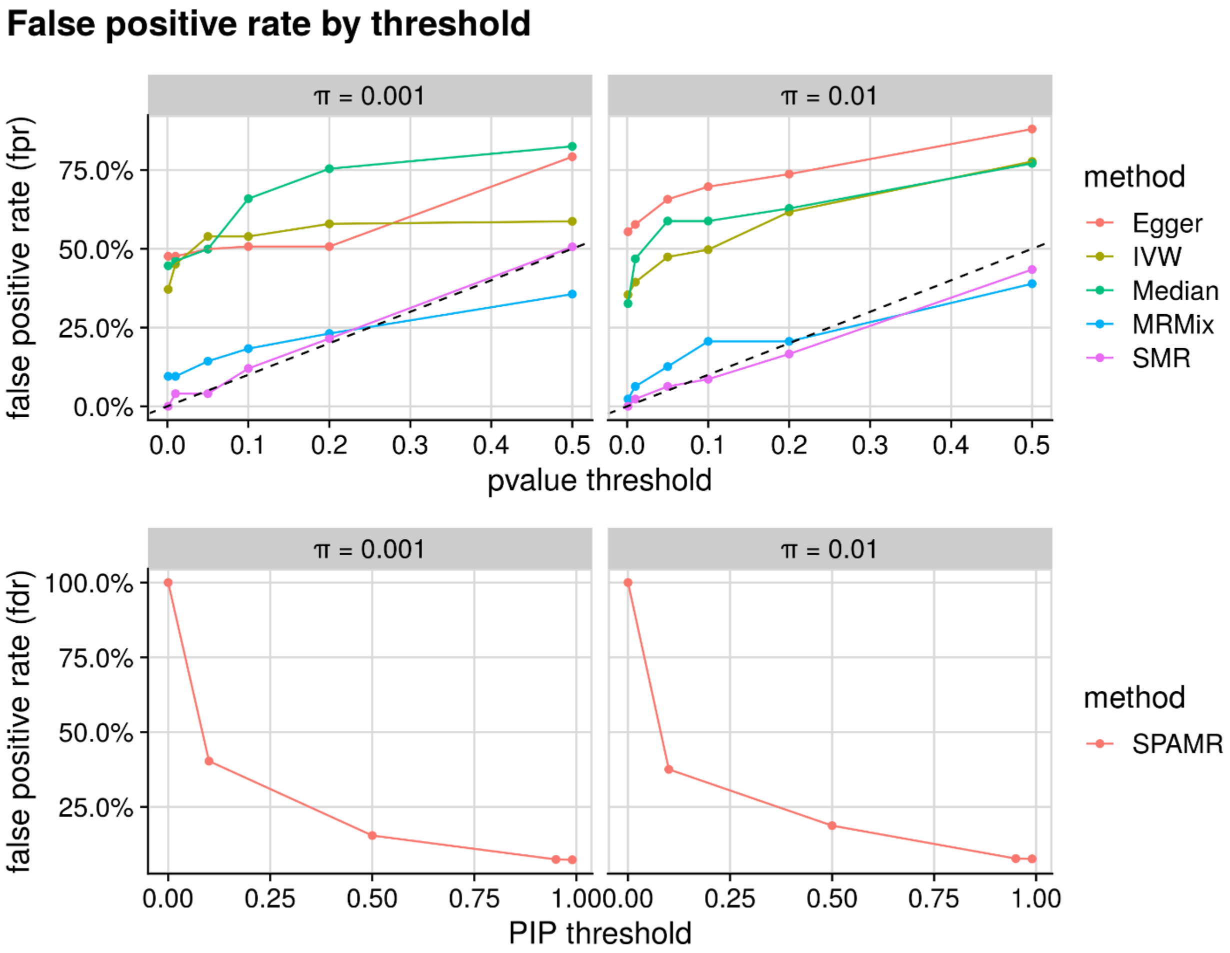
Our results permit several conclusions. Firstly, SPARMR is well-powered to detect causal association in settings where the genetic architecture of the exposure is sparse. As a demonstration of the method, we applied SPARMR to two real-world GWAS from MGI and UK Biobank to estimate the causal association of LDL-linked variation on coronary artery disease. Our method recapitulates a known 'positive control' and demonstrates the ability to characterize the effect of less well studied loci. We expect that our method will prove useful when the assumption of dense causal architecture is unlikely to be true, a setting that appears increasingly common as the ratio of variants to sample sizes continues to increase with the advent of large-scale genotype imputation panels and increased WGS sequencing of biobanks. Traditional least-squares methods without shrinkage priors are challenging by high dimensional variable selection.

We observed several informative results regarding the performance of all the tested methods. SPARMR performed similarly to the weighted median estimator and SMR in terms of power compared to FDR. The weighted median estimator was motivated by the desire apply MR when some unknown proportion of the variants are invalid. In contrast to IVW, the weighted median estimator can provide consistent estimates when at least 50% of the effect size weight comes from valid instruments. This setting has some similarity to ours, when several of the variants that we are testing are not causal due the sparse assumption. Strikingly, the weighted median estimator appears to have similar behavior in this setting with an explicit sparse assumption. SMR also performed well, consistent with its ultra-sparse assumption that only a single variant is causal. It performed remarkably well given that more than a single SNP was assumed to be causal. In practice, choosing the single instrument for SMR remains challenging, though the above results suggest that it performs well even when multiple causal variants are present.

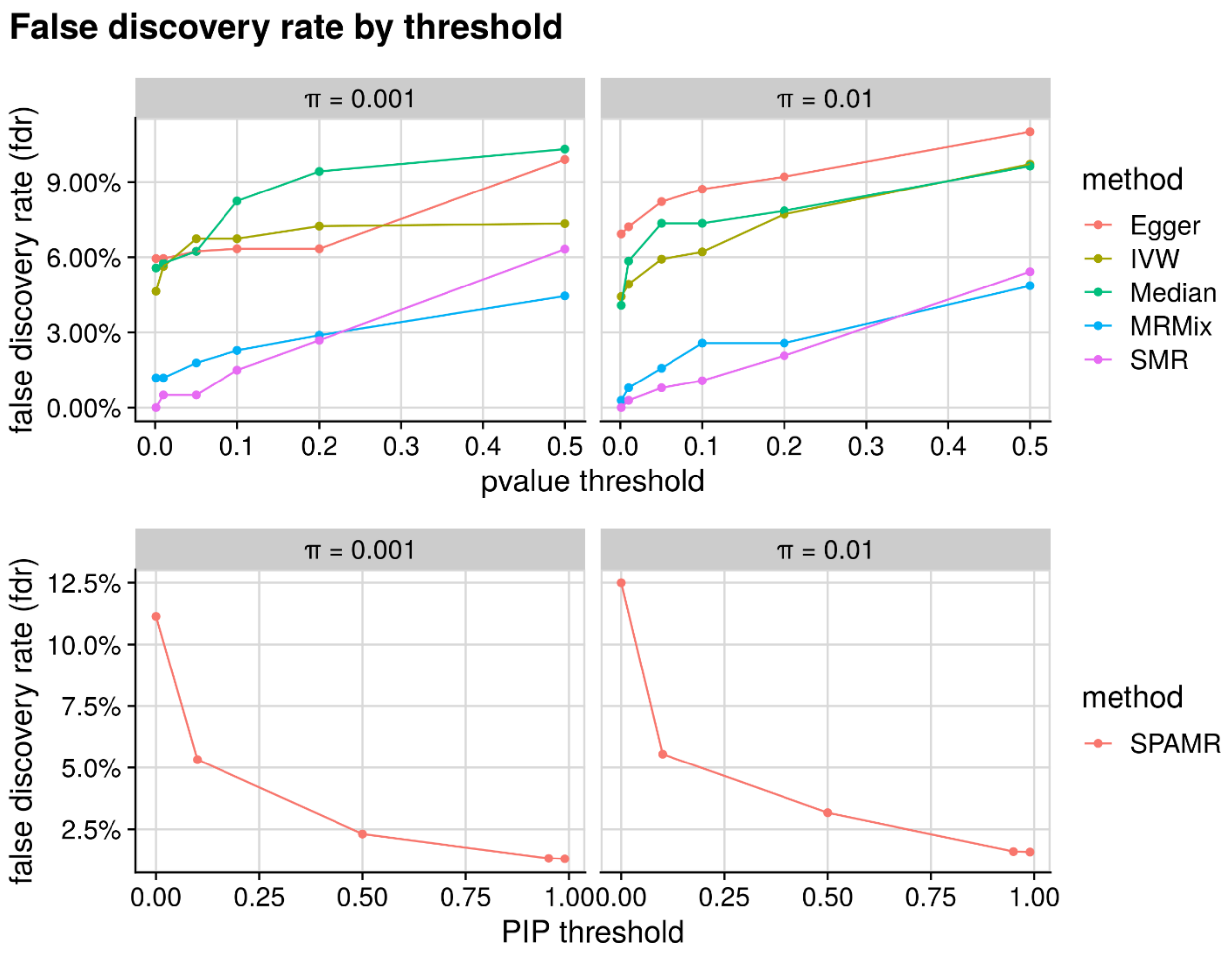
It is challenging to validate the assumptions for MR in practice, which often prohibits the definitive declaration of estimates as causal. The most prominent challenges include instrument selection and horizontal pleiotropy. Our method ameliorates the former in a sparse setting but does not help with the latter. Many recent innovations have focused on detecting and correcting for horizontal pleiotropy. We view these efforts has as essential contributions as pleiotropy is abundant in most MR analyses of complex traits. However, these methods are not without statistical limitations, as introduction of parameters that permit the identification of horizontal pleiotropy results in increased degrees of freedom which reduces power when no pleiotropy is present. Indeed, prior reports have shown that MRMix is underpowered when sample sizes are small relative to traditional approaches (Qi and Chatterjee, 2019). These methods essentially reduce type 1 error at the consequence of increased type 2 error. We propound that study design, especially the choice of exposure and outcome, is the most important determinant of the validity of MR analyses and encourage such analyses in well-understood exposure-outcome systems where knowledge of confounders is known. We posit that SPARMR is best applied in conjunction with principled study design and existing sensitivity checks for validity of instruments, including the potential application of bidirectional MR. Future directions include possible extensions of our method to include a sparse prior for pleiotropic effects, although parameter identifiability would be challenging. Improving the compute time of our method is also a direction of future research – the use of MCMC in this case, although providing principled finite-sample uncertainty estimates, increases the compute time relative to the frequentist methods that use asymptotic approximations to derive the test statistics.



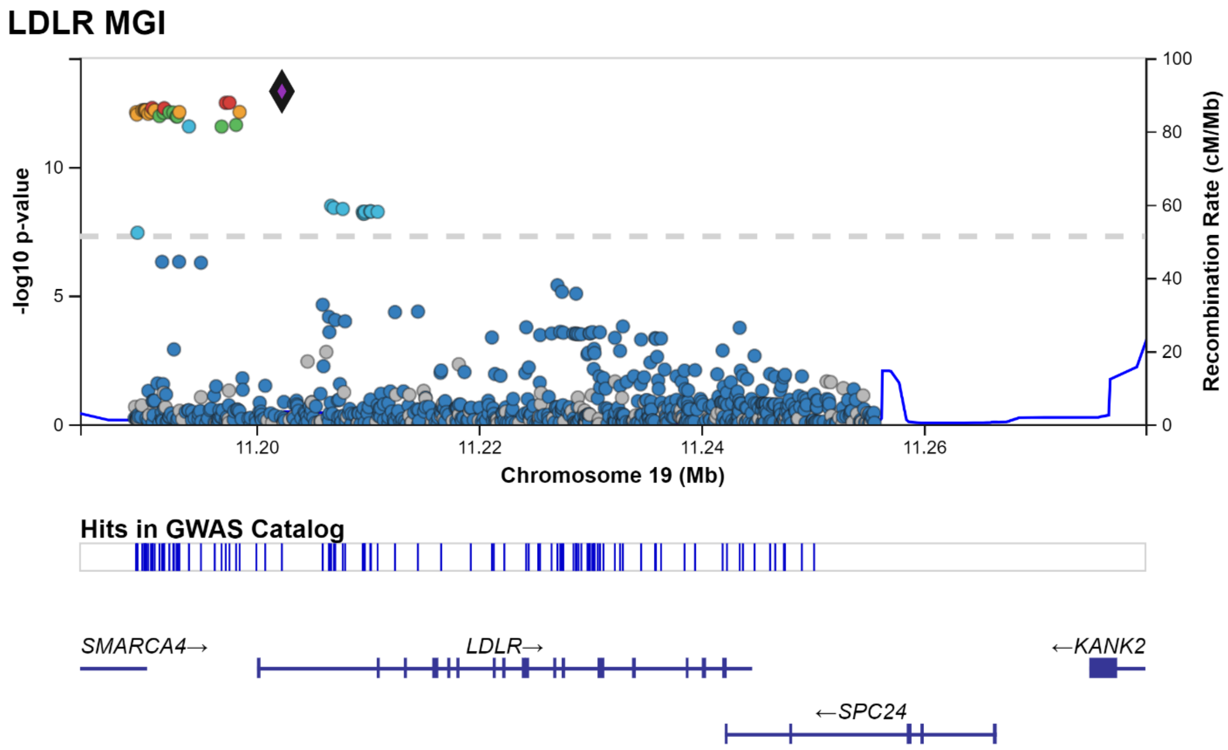
Supplementary Figure 1: The density of the horseshoe distribution with $\tau = 1$. *The distribution has a sharp peak near 0 and has heavy tails to accommodate large effect sizes.*



Supplementary Figure 2: False positive rate by method and test-statistic threshold. In the top panel, the false positive rates of the frequentist methods are displayed for 6 different p-value thresholds (p < .001, .010, .050, .100, .200, .500). The expected coverage of the tests is displayed by the dotted black line which plots y = x. On the left is the setting where the causal variants have a sparsity of .001, and on the right a sparsity of .001. In the bottom panel, the false positive rate of SPARMR is evaluated at 5 PIP thresholds (PIP > 0, .10, .50, .95, .99). The simulations are marginalized over multiple effect size settings.



Supplementary Figure 3: False discovery rate by method and test-statistic threshold. In the top panel, the false discovery rates of the frequentist methods are displayed for 6 different p-value thresholds (p < .001, .010, .050, .100, .200, .500). On the left is the setting where the causal variants have a sparsity of .001, and on the right a sparsity of .001. In the bottom panel, the false positive rate of SPARMR is evaluated at 5 PIP thresholds (PIP > 0, .10, .50, .95, .99). The simulations are marginalized over multiple effect size settings.



Supplementary Figure 3.4: LocusZoom plot of LDL GWAS in the Michigan Genomics Initiative (MGI).

rs6511720 is the highlighted SNP. LD is computed with respected to rs6511720, and is estimated using an 1000 Genomes reference panel.

References

Auton,A. *et al.* (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.

Bates,D. *et al.* (2019) Matrix: Sparse and Dense Matrix Classes and Methods.

Benner,C. *et al.* (2016) FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics*, **32**, 1493–1501.

Berzuini,C. *et al.* (2020) A Bayesian approach to Mendelian randomization with multiple pleiotropic variants. *Biostatistics*, **21**, 86–101.

Bowden,J. *et al.* (2016) Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genetic Epidemiology*, **40**, 304–314.

Bowden,J. *et al.* (2015) Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *International Journal of Epidemiology*, **44**, 512–525.

Boyle,E.A. *et al.* (2017) An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell*, **169**, 1177–1186.

Burgess,S. *et al.* (2013) Mendelian Randomization Analysis With Multiple Genetic Variants Using Summarized Data. *Genet Epidemiol*, **37**, 658–665.

Bycroft,C. *et al.* (2017) Genome-wide genetic data on ~500,000 UK Biobank participants. *bioRxiv*, 166298–166298.

Carvalho,C.M. *et al.* (2009) Handling sparsity via the horseshoe. In, *Artificial Intelligence and Statistics*. PMLR, pp. 73–80.

Carvalho,C.M. *et al.* (2010) The horseshoe estimator for sparse signals. *Biometrika*, **97**, 465–480.

Das,S. *et al.* (2016) Next-generation genotype imputation service and methods. *Nature genetics*, **48**, 1284–1287.

Davey Smith,G. and Ebrahim,S. (2003) 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease?*. *International Journal of Epidemiology*, **32**, 1–22.

Dillon,J.V. *et al.* (2017) TensorFlow Distributions. *arXiv:1711.10604 [cs, stat]*.

Fisher,R.A. (1919) XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Transactions of the Royal Society of Edinburgh*, **52**, 399–433.

Fuchsberger,C. *et al.* (2016) The genetic architecture of type 2 diabetes. *Nature*, **536**, 41–47.

Ge,T. *et al.* (2019) Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nature Communications*, **10**, 1776.

Goldstein,J.A. *et al.* (2020) LabWAS: Novel findings and study design recommendations from a meta-analysis of clinical labs in two independent biobanks. *PLOS Genetics*, **16**, e1009077.

Harris,C.R. *et al.* (2020) Array programming with NumPy. *Nature*, **585**, 357–362.

Hoffman,M.D. and Gelman,A. (2014) The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, **15**, 1593–1623.

Kang,H.M. (2014) EPACTS: efficient and parallelizable association container toolbox.

Kang,H.M. *et al.* (2010) Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, **42**, 348–354.

Kathiresan,S. *et al.* (2008) Polymorphisms Associated with Cholesterol and Risk of Cardiovascular Events. *New England Journal of Medicine*, **358**, 1240–1249.

Klarin,D. *et al.* (2018) Genetics of blood lipids among ~300,000 multi-ethnic participants of the Million Veteran Program. *Nature genetics*, **50**, 1514–1523.

Kumar,R. *et al.* (2019) ArviZ a unified library for exploratory analysis of Bayesian models in Python. *Journal of Open Source Software*, **4**, 1143.

Landrum,M.J. *et al.* (2016) ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Research*, **44**, D862–D868.

Loh,P.-R. (2015) Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics*, **47**, 10.

Mahajan,A. *et al.* (2018) Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nature Genetics*, **50**, 1505–1513.

Morrison,J. *et al.* (2020) Mendelian randomization accounting for correlated and uncorrelated pleiotropic effects using genome-wide summary statistics. *Nature Genetics*, **52**, 740–747.

Novembre,J. and Stephens,M. (2008) Interpreting principal component analyses of spatial population genetic variation. *Nature Genetics*, **40**, 646–649.

Pierce,B.L. and Burgess,S. (2013) Efficient Design for Mendelian Randomization Studies: Subsample and 2-Sample Instrumental Variable Estimators. *American Journal of Epidemiology*, **178**, 1177–1184.

Piironen,J. and Vehtari,A. (2017) On the Hyperprior Choice for the Global Shrinkage Parameter in the Horseshoe Prior. In, Singh,A. and Zhu,J. (eds), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research. PMLR, Fort Lauderdale, FL, USA, pp. 905–913.

Price,A.L. *et al.* (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, **38**, 904–909.

Pritchard,J.K. (2001) Are Rare Variants Responsible for Susceptibility to Complex Diseases? *Am J Hum Genet*, **69**, 124–137.

Qi,G. and Chatterjee,N. (2019) Mendelian randomization analysis using mixture models for robust and efficient estimation of causal effects. *Nature Communications*, **10**, 1941.

Ripke,S. *et al.* (2014) Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, **511**, 421–427.

Sinnott-Armstrong,N. *et al.* (2021) Genetics of 35 blood and urine biomarkers in the UK Biobank. *Nature Genetics*, **53**, 185–194.

Taliun,D. *et al.* (2021) Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature*, **590**, 290–299.

Teslovich,T.M. *et al.* (2010) Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, **466**, 707–713.

Visscher,P.M. *et al.* (2017) 10 Years of GWAS Discovery: Biology, Function, and Translation. *The American Journal of Human Genetics*, **101**, 5–22.

Wainschtein,P. *et al.* (2019) Recovery of trait heritability from whole genome sequence data. *bioRxiv*, 588020.

Wen,X. *et al.* (2016) Efficient Integrative Multi-SNP Association Analysis via Deterministic Approximation of Posteriors. *The American Journal of Human Genetics*, **98**, 1114–1129.

Willer,C.J. *et al.* (2013) Discovery and refinement of loci associated with lipid levels. *Nature genetics*, **45**, 1274–1283.

Yang,J. *et al.* (2012) Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nature Genetics*, **44**, 369–375.

Yavorska,O.O. and Burgess,S. (2017) MendelianRandomization: an R package for performing Mendelian randomization analyses using summarized data. *International Journal of Epidemiology*, **46**, 1734–1739.

Yengo,L. *et al.* (2018) Meta-analysis of genome-wide association studies for height and body mass index in ∼700000 individuals of European ancestry. *Hum Mol Genet*, **27**, 3641–3649.

Yuan,Z. *et al.* (2020) Testing and controlling for horizontal pleiotropy with probabilistic Mendelian randomization in transcriptome-wide association studies. *Nature Communications*, **11**, 3861.

Zhang,F. and Lupski,J.R. (2015) Non-coding genetic variants in human disease. *Hum Mol Genet*, **24**, R102-110.

Zhou,W. *et al.* (2018) Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature Genetics*, **50**, 1335–1341.

Zhou,X. *et al.* (2013) Polygenic Modeling with Bayesian Sparse Linear Mixed Models. *PLOS Genetics*, **9**, e1003264.

Zhu,X. and Stephens,M. (2017) BAYESIAN LARGE-SCALE MULTIPLE REGRESSION WITH SUMMARY STATISTICS FROM GENOME-WIDE ASSOCIATION STUDIES. *Ann Appl Stat*, **11**, 1561–1592.

Zhu,Z. *et al.* (2016) Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature Genetics*, **48**, 481–487.

Zuk,O. *et al.* (2014) Searching for missing heritability: Designing rare variant association studies. *PNAS*, **111**, E455–E464.

Chapter 4 Clonal Hematopoiesis is Driven by Aberrant Activation of TCL1A

Main

Aging is characterized by the accumulation of somatic mutations which frequently are benign. However, some mutations confer proliferative advantages resulting in an expanded lineage of cells, termed a clone. Clonal hematopoiesis of indeterminate potential (CHIP) is defined by the acquisition of specific cancer-associated mutations in hematopoietic stem cells (HSC) from persons without a blood cancer(Steensma *et al.*, 2015). Previous reports have associated CHIP with increased risk for hematologic malignancy, coronary heart disease, and mortality(Jaiswal *et al.*, 2014; Genovese *et al.*, 2014; Xie *et al.*, 2014). In contrast to small clones, which are ubiquitous in older individuals and are benign, large clones are less common and more likely to result in hematologic malignancy and cardiovascular disease(Jaiswal *et al.*, 2017; Bick Alexander G. *et al.*, 2020). However, few determinants of clonal expansion have been identified to date, partially due to the lack of large cohorts with serially sampled blood over several years. Using 5,551 CHIP carriers derived from 127,946 deep (38x) whole genomes from the NHLBI Trans-omics for Precision Medicine (TOPMed) initiative(Taliun *et al.*, 2019; Bick *et al.*, 2020), here we developed a sequencing based method for approximating the rate of clonal expansion from a single timepoint, termed PACER, which was validated using ultra-high depth

(>300x) longitudinal sequencing. Using PACER, we performed the first large-scale investigation of the germline determinants of clonal expansion.

Development and validation of PACER

We identified high-confidence somatic mutations in peripheral blood by analyzing the TOPMed WGS with GATK Mutect2(Cibulskis *et al.*, 2013). To remove sequencing artifacts and germline variants we performed stringent variant filtering and quality control. We identified CHIP carriers using a curated list of leukemogenic driver mutations (Supplementary Table 1) (Methods). We identified 6,158 CHIP mutations in 5,551 individuals. As shown in our previous report, the prevalence of CHIP was strongly associated with age at blood draw, and >75% of these mutations were in *DNMT3A*, *TET2*, or *ASXL1*.

We hypothesized that the number of passenger mutations present in the WGS data of CHIP carriers could be used to estimate the rate of expansion of the clone. As the passengers accrue at a rate that is fairly constant rate over time and that is similar between individuals(Osorio *et al.*, 2018), they can be used to date the acquisition of the driver (Figure 1a). For two individuals of the same age and with clones of the same size, we expect the clone with more passengers to be more fit, as it expanded to the same size in less time. Typically, one would need to isolate single-cell colonies derived from HSCs in order to calculate the total passenger mutation burden (Lee-Six *et al.*, 2018). However, we hypothesized that this measure could also be approximated from WGS data from whole blood DNA. The variant allele fraction (VAF), defined as the proportion of sequencing reads at a locus containing the mutant allele, is an approximate measure of clone size. As the clone expands, the VAF of both the driver and passenger mutations increases. The number of passengers in any given cell is simply the sum of the mutations present prior to the acquisition of the driver event (founding passengers) and mutations acquired after the driver event (subclonal passengers). At VAF values of greater than 5-10%, the detectable passengers are far more likely to be founding passengers than subclonal passengers. This is because the subclonal passengers are private to each subsequent division of the original mutant cell, and, in the absence of second driver event, quickly fall below the limit of detection in bulk tissue. Furthermore, as the size of the clone also determines the number of detectable passengers from WGS, high fitness clones will harbor more detectable passengers than lower fitness clones that arose at the same time. Based on these observations, we used the detectable passengers as a composite measure of clone fitness and birth date.

To estimate the number of passengers, we first obtained Mutect2 variant calls from the whole genome for each CHIP carrier and a subset of people without detectable CHIP. As the raw variant calls are expected to contain a combination of true somatic variants, germline variants, and sequencing artifacts, we implemented a series of filters to enrich for the detection of true passengers. As the frequency of mutations across individuals is a function of both mutation rate and fitness(Watson *et al.*, 2020), we first selected only those variants that were found in a single individual (singletons) in the dataset, as recurrent variants are enriched for germline polymorphisms and recurrent artifacts. We also excluded variants with a VAF greater than 35%, as these would be enriched for germline polymorphisms. As different base substitutions varied in their association with age at blood draw, we selected only C-T and T-C mutations, as these were the most strongly age-associated. On average, the CHIP carriers had 237 passengers (95% CI: 229-246), the median value was 206, and the maximum value was 16,279. Of the CHIP carriers, 90% had a single driver mutation. The passengers were enriched by 54% (95% CI: 51%-57%) in the CHIP carriers (Extended Data Fig 1) compared to the controls after adjusting for age and study using a negative binomial regression. In the controls without CHIP, we presumed the detected “passengers” were incompletely removed artifacts or, in some people, reflective of unidentified clonal hematopoiesis. The passengers were also positively associated with age, on average increasing by 13.7% (95% CI: 13.0%-14.3%) each decade.

We validated the passengers as an estimator of fitness theoretically and empirically. For the theoretical validation, we constructed a simulation of HSC dynamics to characterize the relationship between fitness and detectable passenger counts (Supplementary Information 1). We estimated a passenger mutation rate per diploid genome per year of 2.3, or a per-base pair rate of $3.83x10^{-10}$. Assuming 100,000 HSCs, this results in a per-base-pair passenger mutation rate of $3.83x10^{-15}$ per HSC clone per year without correction for the sensitivity of the sequencing technology used. This number is substantially lower than previous estimates using WGS from single hematopoietic colonies, likely due the low sensitivity of detecting true passengers in whole blood DNA compared to the gold standard of single-cell derived colonies. Nonetheless, we were able to use this data to derive a hierarchical Bayesian latent-variable estimator of clone fitness (Methods) and confirmed its strong correspondence to the observed passenger counts (supplement S2).

To empirically validate the passengers, we used ultra-high depth sequencing in 80 CHIP carriers from the Women’s Health Initiative (WHI) from two time points using single-molecule molecular inverse probe sequencing(Hiatt *et al.*, 2013) (smMIPS) targeted to the CHIP driver genes (Methods). We called somatic variants in these samples using an ensemble of VarScan(Koboldt *et al.*, 2012), GATK Mutect2(Cibulskis *et al.*, 2013), and manual inspection through IGV(Robinson *et al.*, 2011) (Methods). We defined clonal expansion by dividing the change in VAF by the change in time (years) $\frac{dVAF}{dT}$ of the driver variants identified at the first blood draw. Of the sequenced carriers, the majority had clones that were constant in size or expanded. We constructed a simple estimator of $\frac{dVAF}{dT}$ using only the passengers, VAF, and age from the first blood draw (Methods). This estimator predicted the inverse normal transformed $\frac{dVAF}{dT}$ (Rsq = 32.5%, Adjusted Rsq = 28.6%, pvalue = 1.5e-4, Figure 1b, Figure 1c). After adjusting for the passenger counts, age was negatively associated with $\frac{dVAF}{dT}$, suggesting that clones acquired later in life were on average less fit than those acquired by younger individuals. We also observed that VAF at the first time point was negatively associated with $\frac{dVAF}{dT}$ after adjustment for the other covariates, which may reflect the largest clones saturating in clonality. These results suggested that inferring clonal expansion from age and VAF-adjusted passenger mutation counts was able to predict not only past growth, but also future growth rate. We termed this approach PACER (passenger-approximated clonal expansion rate).

PACER predicts fitness of distinct driver mutations

Building on recent computational estimates of variant fitness(Watson *et al.*, 2020), we estimated the distribution of passengers across the most common CHIP driver genes. We stratified the *DNMT3A* carriers by whether the driver mutation was a missense mutation at position 882 into *DNMT3A* R882+ and *DNMT3A* R882- carriers. We used *DNMT3A* R882- as a reference point and estimated the relative abundances of passengers in other genes using negative binomial regression adjusting for age and study. Consistent with previous reports, splicing genes (*SF3B1*, *SRSF2*, *U2AF1*) and *JAK2* V617F mutations had the highest PACER values, while *DNMT3A* R882- was among the lowest (Figure 1d). Mutations in *TET2*, *ASXL1*, *PPM1D*, *TP53*, *ZBTB33*, and *GNB1* were in the next tier and had approximately the same level of fitness estimated from PACER. Relative to the R882- carriers, we observed a modest increase fitness in the R882+ carriers. These observations are concordant with prior empirical estimates of variant fitness derived from longitudinal sequencing of samples with clonal hematopoiesis(Desai *et al.*, 2018).

Genome wide association study of PACER

To characterize the molecular pathways associated with clonal expansion, we performed a genome-wide association study (GWAS) of the inverse normal transformed passenger counts of CHIP carriers. We included age at blood draw, study, VAF, and the first ten genetic ancestry principal components, and used SAIGE(Zhou *et al.*, 2018) to estimate the single variant association statistics among 19,913,304 variants. The GWAS identified a single locus at genome-wide significance at *TCL1A* (Figure 2a). We used SuSIE(Wang *et al.*) to fine-map (Methods) a 200kb region surrounding *TCL1A* which identified a credible set containing a single variant rs2887399 (Extended Data Fig. 2). Each additional T allele was associated with a decrease in passenger count z-score by 0.15 (pvalue = 4.5e-12). The alt-allele is common, occurring in 26% of TOPMed haplotypes. rs2887399 lies in a core promoter of *TCL1A* as defined by the Ensembl regulatory build(Dr *et al.*, 2015) (Figure 3a) 162 base-pairs from the canonical transcription start site (TSS) and in a CpG island. Analysis of the variant by the Open Targets(Carvalho-Silva *et al.*, 2019) variant-to-gene (V2G) function also nominated *TCL1A* as the causal gene. *TCL1A* has been implicated in prior reports as driver gene in lymphocytic malignancy, but no connection to clonal hematopoiesis or HSC biology has previously been described.

We then asked whether any genetic variation associated with the passenger counts was specific to different CHIP mutations. We performed separate GWAS of passenger counts for carriers of *TET2*, *DNMT3A*, *ASXL1*, and splicing mutations. In *TET2* carriers, we observed variation at the *SASH1-UST* locus was associated with passenger counts. The lead variant rs4897025 is a common (MAF = 43%) intergenic variant that was associated with decreased passenger count burden (beta = -0.3, pvalue = 2.8e-8). Previous reports have observed that downregulation of *SASH1* is associated with increased risk for breast cancer(Zeller *et al.*, 2003). In *DNMT3A* carriers we observed no association between rs4897025 and passenger counts (pvalue = 7.4e-1), consistent with its effect in the non-stratified passenger count GWAS (pvalue = 1.4e-2). In *TET2* carriers the effect size of alt-alleles of rs2887399 was larger than in the non-stratified GWAS (beta = -0.15 in non-stratified GWAS, beta = -0.24 in *TET2* carrier GWAS). We observed no other germline variation that was associated with passenger counts in the other CHIP gene stratified GWAS, possibly due to limiting sample size.

We examined the association between the burden of rare variation with passenger counts in the 200kb region surrounding *TCL1A*. We used the SCANG rare variant scan procedure(Li *et al.*, 2019) to estimate the association, including all variants with a MAC <= 300 (MAF <= 3.7%). The SCANG procedure estimates the association between rare variants in moving windows across the genome and estimates the size of the windows. SCANG did not identify any regions at exome-wide significance (2.5e-06), though did identify one region within an order of magnitude (pvalue = 6.6e-06, family-wise pvalue = 2e-03, Extended Data Fig. 3). After conditioning on the rs2887399 genotypes in the rare variant analysis, the signal was attenuated, suggesting limited evidence for an independent rare-variant signal from rs2887399 in the same region (<1 Mb, Extended Data Fig. 4). We identified only 10 putative loss-of-function (pLOF) carriers of TCL1A TOPMed wide, so were underpowered to examine the burden of these variants.

We performed an expanded search of rare variation associated with the passengers. We used 1,698 genes associated with 'cancer' according to Open Targets (Carvalho-Silva *et al.*, 2019) to define variant groups (Supplementary Table 2). We performed SCANG association tests at every gene and its 150 Kb flanking region, including both coding and non-coding variants with a MAC <= 300. We identified 15 windows associated with passenger counts at Bonferroni significance (pvalue = 2.9e-5, Supplementary Table 3). We identified an intergenic region 113kb from the TSS of *TNFAIP3* (pvalue = 5.4e-7) that is a distal enhancer of TNFAIP3 (GeneHancer(Fishilevich *et al.*, 2017)).

As the allele frequency of rs2887399 varies by population, we asked whether passenger count was associated with the first two genetic ancestry principal components. We observed a positive association between values on the PC1 axis with singleton counts. Even after conditioning on the rs2887399 dosage, the association remained. A linear regression with rs2887399 dosage and the first two principal components as covariates explained 4% of the variation in the inverse-normal transformed passenger counts. Ancestry estimation using RFMix(Maples *et al.*, 2013) indicated a modest depletion of passengers in Sub-Saharan African genomes relative to European and East Asian genomes (Methods, Extended Data Fig. 5).

Association of TCL1A genotype to CHIP driver genes

We asked whether the association between rs2887399 and passenger counts was modified by CHIP driver gene. Using *DNMT3A* as the reference, we investigated whether other genes had different effect estimates for rs2887399. We observed that alt-allele dosage in rs2887399 was more protective in *TET2* than *DNMT3A* (beta = -0.23, pvalue = 2e-03, Figure 2b), but we were underpowered to detect effects in other genes. These results suggest that the protective effects of rs2887399 vary by CHIP driver mutation and are weaker in *DNMT3A* compared to *TET2*. As the alt-homozygotes of rs2887399 were depleted for other CHIP mutations, we were underpowered to estimate the association between rs2887399 dosage and passenger counts in the other CHIP genes.

Functional impact of TCL1A genotype on CHIP

To further interrogate the effect of rs2887399 on CHIP, we also performed association tests between the variant and the prevalence of specific driver genes. In our previous analysis(Bick *et al.*, 2020) we reported that the T allele was associated with increased risk for DNMT3A mutation. Here, in an expanded analysis of 74,974 individuals, we observed that rs2887399 is protective for (Figure 2c) multiple non-*DNMT3A* driver mutations, and splicing mutations. The alternate homozygous genotype was associated with decreased risk of acquisition of multiple non-*DNMT3A* mutations (OR = 0.20, 95% CI: 0.06 – 0.51, Figure 3d, Methods). These results indicate that rs2887399 increases risk for low fitness *DNMT3A* clones but is protective against clones that more strongly predict progression to frank hematologic malignancy(Desai *et al.*, 2018), including *JAK2*, *ASXL1*, *SRSF2*, and *SF3B1*, and was especially protective against the acquisition of >1 non-*DNMT3A* driver mutations. The latter are particularly relevant clinically, as these persons have the greatest risk of transformation, and in some cases may already have early-stage MDS.

Previous analysis of blood cell indices in UK Biobank(Bycroft *et al.*, 2017) have implicated rs2887399 in reduced blood cell counts (Figure 2d), consistent with altered hematopoiesis. To further characterize the disease associations of rs2887399, we performed a phenome-wide association study (PheWAS) lookup in UK Biobank. Although no genome-wide significant associations were identified among the case-control phenotypes, the alt-allele was nominally protective against myeloproliferative neoplasms (beta = -0.12, pvalue = 2.7e-02) and leukemia (beta = -0.11, pvalue = 1.0e-02). Previous reports have also identified that the alt-allele of rs2887399 increases risk for mosaic loss of the Y chromosome(Thompson *et al.*, 2019; Zhou *et al.*, 2016) (beta = 0.20, pvalue = 6.0e-11), indicating a convergence of variation at the locus affecting multiple distinct clonal phenomena. A PheWAS lookup of gene-based test statistics using 45,596 UK Biobank(Bycroft *et al.*, 2017) exomes identified a nominal association between *TCL1A* coding variants with other anemias (UKB exome phewas(Zhao *et al.*, 2020), phecode 285, pvalue = 2.3e-2).

Next, we functionally characterized the rs2887399 locus. We first asked if the variant was associated with *TCL1A* expression in any cell type. As identified in the GTEx v8 eQTL release(Consortium, 2020), the alt-allele reduces expression of *TCL1A* in whole blood (normalized effect size = -0.13, pvalue = 1.4e-5, Figure 3a). The association is likely driven by B-cells, as *TCL1A* is highly expressed in B-cells but appears to have absent or low expression in all other cell types except plasmacytoid dendritic cells. We also did not see expression of *TCL1A* in normal human HSCs or myeloid progenitors in publicly available gene expression datasets. We next asked whether CHIP-associated mutations might alter the regulation of the *TCL1A* locus in HSCs.

Given the mutation specific associations of rs2883799, we asked whether the regulation of *TCL1A* varied by CHIP driver gene mutation. Using a reference of chromatin accessibility in normal and pre-leukemic HSCs (pHSCs)(Corces *et al.*, 2016), we examined the ATAC-seq readout at the *TCL1A* promotor. Consistent with the lack of *TCL1A* transcripts in normal HSCs, we observed that the promoter was not accessible in either normal human donor HSCs, or pHSCs from patients with AML without any driver mutations. We also did not observe accessible chromatin in two carriers of *DNMT3A* mutated pHSCs. In contrast, in the two patients with *TET2* mutated pHSCs the *TCL1A* promoter was clearly accessible. These observations led us to propose the following mechanistic model: Normally, the *TCL1A* promoter is inaccessible and gene expression is absent in HSCs. In the presence of driver mutations in *TET2*, *ASXL1*, *SF3B1*, *SRSF2*, *JAK2*, and possibly other genes, the *TCL1A* promoter becomes accessible, permitting gene expression and driving clonal expansion of the mutated cells. The presence of the alt-allele of rs2887399 inhibits accessibility of chromatin at the *TCL1A* promoter, leading to reduced expression of *TCL1A* RNA, and abrogating clonal advantage due to the mutations.

DISCUSSION

Using the largest dataset of CHIP whole genomes to date, we have generated several novel insights into clonal expansion. We developed PACER, a method that allows us to infer clonal expansion rate from a single time point. The passenger counts represent a composite measure of the fitness and birth date of an underlying clone and provides a simple predictor of clonal expansion. Our results extend and apply recently developed theory on the evolutionary fitness of clones to permit estimation of fitness of a clone within a single individual, but used passenger mutation counts in contrast to previous efforts which used the VAFs of driver variants. Unlike prior methods, PACER can also be used to perform association tests for novel factors associated with clonal expansion. Using PACER we show that the fitness effect of mutations in different driver genes can vary considerably, in accordance with other recent reports.

In a GWAS for PACER, we identified as the top hit a common variant of large effect in the promoter of *TCL1A*. Remarkably, this variant is associated with protection from several CHIP driver variants, including gene mutations that heretofore have not had known targets promoting clonal expansion such as *TET2*, *ASXL1*, *SF3B1*, and *SRSF2*. Analysis of a chromatin accessibility atlas(Corces *et al.*, 2016) nominated a putative mechanism where some CHIP mutations allow chromatin to become accessible at the *TCL1A* promoter and the gene to be transcribed, an effect which may be abrogated by the protective allele. The large protective effect seen with the rs2887399 alt-allele suggests that *TCL1A* expression may be the dominant factor for clonal expansion due to these mutations. Prior work has shown that it is deregulated in T-cell leukemia and lymphoma(Hecht *et al.*, 1984), chronic lymphocytic leukemia, and that it is a co-activator of Akt kinases(Laine *et al.*, 2000), but there have not been prior studies linking *TCL1A* to HSC biology or clonal hematopoiesis. How *TCL1A* expression causes clonal expansion of HSCs is an important question for future studies.

PACER represents a powerful tool for studying factors influencing clonal expansion, but our study has limitations. The sequencing coverage in TOPMed WGS was 38x, which inhibits detection of mutations with VAFs below 5%(Bick *et al.*, 2020). More sensitive assays would increase detection of both driver and passenger mutations, would allow for phylogenetic analyses of the mutations that require accurate variant allele fractions and would reduce error in estimation of clonal expansion using PACER Additionally, the use of bulk-WGS precludes analysis of the co-occurrence of passenger and driver mutations in the same HSCs, which would refine our definition of passenger mutations.

Nonetheless, PACER represents a powerful tool for studying clonal expansion in human WGS datasets. Despite the limitations of the sequencing technologies used, we also briefly remark that analysis of high-VAF passengers is key to the estimation of clonal expansion with this method, as it is only the passengers that occur on the predominant clone that is informative here. Although our analysis has focused on clonal expansion in blood cells, PACER may be adapted to study clonal expansion in any tissue where pre-malignant clones exist.

Methods

Study Samples

Whole genome sequencing (WGS) was performed on 127,946 samples as part of 51 studies contributing to Freeze 8 NHLBI TOPMed program as previously described(Taliun *et al.*, 2019). None of the TOPMed studies included selected individuals for sequencing because of hematologic malignancy. Each of the included studies provided informed consent. Age was obtained for 82,807 of the samples, and the median age was 55, the mean age 52.5, and the maximum age 98. The samples have diverse reported ethnicity (40% European, 32% African, 16% Hispanic/Latino, 10% Asian).

WGS Processing, Variant Calling and CHIP annotation

BAM files were remapped and harmonized through the functionally equivalent pipeline(Regier *et al.*, 2018). SNPs and indels were discovered across TOPMed and were jointly genotyped across samples using the GotCloud pipeline(Jun *et al.*, 2015). An SVM filter was trained to discriminate between high- and low-quality variants. Variants were annotated with snpEff 4.3(Cingolani *et al.*, 2012). Sample quality was assessed through mendelian discordance, contamination estimates, sequencing converge, and among other quality control metrics.

Putative somatic SNPs were called with GATK Mutect2(Cibulskis *et al.*, 2013), which searches for sites where there is evidence for alt-reads that support evidence for variation, and then performs local haplotype assembly. We used a panel of normals to filter sequencing artifacts and used an external reference of germline variants to exclude germline calls. We deployed this pipeline on Google Cloud using Cromwell(Voss *et al.*, 2017).

As described in our previous report, samples were annotated as having CHIP if the Mutect2 output contained at least one variant in a curated list of leukemogenic driver mutations with at least three alt-reads supporting the call. We expanded the list of driver mutations to include those in recently identified CHIP genes, increasing the number of CHIP cases from our previous report.

We called somatic singletons by identifying somatic variants that appeared in a single individual among the CHIP carriers and 23,320 additional controls for a total of 28,391 individuals. We excluded any variant that appeared in TOPMed Freeze 5 (463 million variants). We excluded variants with a depth below 25 or above 100 and excluded any variants in low complexity regions or segmental duplications, as these are challenging for variant calling. We only included somatic singletons that were aligned to the primary chromosomal contigs. We excluded any variant with a VAF exceeding 35% as these may be enriched for germline variants that were not included in our other filters. We used cyvcf2(Pedersen and Quinlan, 2017) to parse the Mutect2 VCFs and encoded each variant in an int64 value using the variant key encoding(VariantKey: A Reversible Numerical Representation of Human Genetic Variants | bioRxiv). We developed a bespoke Python application to perform the singleton identification and filtering.

Amplicon sequencing validation

Targeted sequencing of the CHIP driver genes from 80 samples from the Women's Health Initiative (WHI) was performed using single-molecule molecular inversion probe sequencing (smMIPS(Hiatt *et al.*, 2013)). Reads were aligned with bwa-mem and processed with the mimips pileline (cite). We called somatic variants using an ensemble of VarScan(Koboldt *et al.*, 2012), Mutect2(Cibulskis *et al.*, 2013), and manual inspection with IGV(Robinson *et al.*, 2011).

Single Variant Association

Single variant association for each variant in Freeze 8 with a MAC > 20 was performed with SAIGE(Zhou *et al.*, 2018) using the TOPMed Encore analysis server. To identify associations between rs2887399 and the acquisition of specific CHIP mutations, we used the same methods as our previous report on an analysis set of 74,974 individuals, including 4,697 cases and 70,277 controls. Age, genotype inferred sex, the first ten genetic ancestry principal components, and study were included as covariates.

We performed SAIGE single variant association analyses on the passengers including age at blood draw, sex, VAF, study, and the first ten genetic ancestry principal components as covariates. We applied an inverse normal transformation to the passenger counts. We declared variants from this analysis as significant if their pvalue was less than 5e-8.

Estimation of association between rs2887399 genotypes and CHIP mutation acquisition

We coded the rs2887399 genotypes as a categorical variable rather than a linear quantitative coding to estimate effects separately for the heterozygotes and the alt-homozygotes using the ref-homozygotes as the reference level. We estimated the associations using firth logistic regression to reduce bias in estimation resulting from low cell counts(Ma *et al.*, 2013), and included age, genotype inferred sex, and the first ten genetic ancestry components as covariates.

Finemapping of the TCL1A region

We applied SuSIE(Wang *et al.*) to the genotypes included in a 200kb region surrounding TCL1A. We used the same covariates as the single variant association analysis. We used the posterior inclusion probabilities (PIP) and credible sets identified by SuSIE to identify the putative causal variant. We used LD directly calculated on the genotypes as opposed to an external reference.

Rare Variant Analyses

We performed gene-based tests on 1,698 cancer associated genes their flanking regions using the SCANG(Li *et al.*, 2019) procedure. We identified these genes by downloading the targets associated with cancer in Open Targets(Carvalho-Silva *et al.*, 2019), and then filtered to include only genes with an association score of 1.0. The most prevalent CHIP driver genes were included among this list. We used the inverse normal transformed passenger counts as the phenotype with the same covariates as before. We specified the minimum size of the grouped regions as 30 variants and the maximum as 200. We included all PASS variants with a minor allele count greater than four and less than 300 (MAF of 3.7% in the analyzed samples). We parsed the genotypes using cyvcf2 and stored them as dgCMatrix using the Matrix(Bates *et al.*, 2019) package from the R 3.6.1 programming language(R Core Team, 2020).

We set the p-value filter to calculate SKAT test-statistics at 5e-4. We did not group the variants by annotation and we declared regions as significant if their pvalue was less than 2.9e-5 (.05 / 1,698). We controlled for relatedness by incorporating a sparse kinship matrix as estimated by the PC-AiR method from the GENESIS R package(Gogarten *et al.*, 2019). We specified separate residual variance terms for each study to control for heterogeneous residual variance. We grouped together all studies where the number of analyzed samples was less than 200.

Enrichment of passengers by driver gene

We estimated the association between the driver genes and the passenger counts using DNMT3A as the reference in a negative binomial regression using the glm.nb function from the MASS R package(Venables and Ripley, 2002). We included age, study, VAF, and sex as covariates. We included driver genes with at least 30 mutations and reported genes that had a different effect relative effect than DNMT3A if the pvalue of the coefficient was less than 1e-2.

Estimation of passenger mutation rate, clone fitness, and clone birth date

We developed a hierarchical Bayesian latent variable model using the Stan(Stan Development Team, 2020b, 2020a) probabilistic programming language. We used the negative binomial likelihood with a mean and overdispersion parameterization to facilitate interpretation. We used the identity function to link the passenger counts to the predictors as we modeled the effects on an additive scale. We modeled the expectation and overdispersion of the passenger counts observed at time (t_i) as

$$\mathrm{E}\big(counts_i(t_i)\big) = \mu T_i + s_i(t_i \ - T_i) + \ \alpha_k$$

$$counts_i(t_i) \sim NB(E(counts_i(t_i)), I(i \in CHIP)\theta_0 + \big(1 - I(i \in CHIP)\big)\theta_1)$$

Where T_i is the time of the driver acquisition for sample *i* with a blood draw at time t_i, μ is the mutation rate per diploid genome per year for the HSC population, s_i is the fitness of the clone, and α_k represents a study specific random intercept for sample *i* included in study *k.* We can interpret $t_i \ - T_i$ as the lifetime of the clone in years. We used a negative binomial likelihood as there was overdispersion relative to a Poisson distribution.

We included several constraints and priors on the parameters to make them identifiable. We constrained T_i to be positive but exceeded by t_i such that the parameter would be in yearly units. We included case-control specific overdispersion terms θ_0 and θ_1 as the CHIP carriers had greater dispersion. To adjust for batch effects, we included a random intercept, as the amount of singletons in controls varied by study.

To include the constraint on T_i, we defined $T_i = \psi_i * age_i$, with ψ_i constrained between 0 and 1, and age_i is the age at blood draw. We placed a Beta(1, 1.3) prior on ψ_i, which is equivalent to the supposition that the driver mutation is twice as likely to be acquired in the second half of life (at the time of blood draw) then the first. We assumed the study specific deviations were exchangeable with respect to a $N(0,20)$ prior, providing some shrinkage on the study specific intercepts. We placed a $N(0,1)$ prior on the s_i parameter to aid identification. Further details are described in the supplement.

To estimate the posterior, we used the Stan Hamiltonian Monte-Carlo (HMC) sampler with four separate chains, and used 400 samples of burn-in. We assessed convergence using the Rhat and effect-sample size statistics. We tried multiple parameterizations to reduce the number of divergent transitions. We performed posterior predictive checks to assess the model fit.

Simulation of HSC dynamics

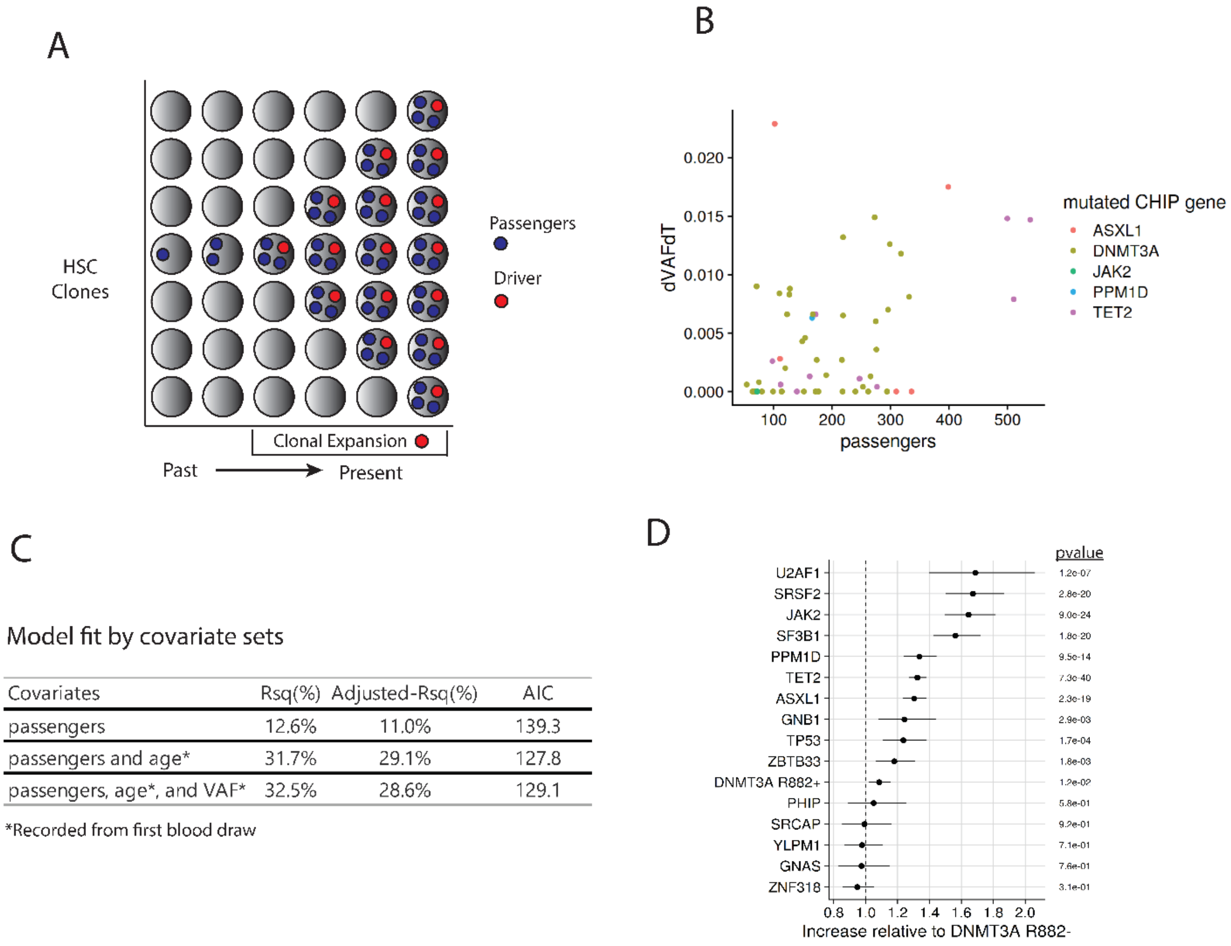
We simulated the number of cells within an HSC clone as a birth-death continuous time Markov chain, which models the size of an HSC clone as the composite of simultaneous Poisson birth and Poisson death point processes. Following Watson et al.(Watson *et al.*, 2020), HSCs could transition to one of three states: asymmetric renewal, symmetric self-renewal, and symmetric differentiation. The rate of transition was determined by the symmetric differentiation rate of the cell per year, which was set to five. The symmetric self-renewal and symmetric differentiation increase and decrease the size of the HSC clone respectively. As asymmetric division does not affect the size of the clone, we did not explicitly simulate transition to this state. The proclivity towards self-renewal was determined by the fitness of the clone. We set the entire HSC population to acquire a single driver mutation during the ‘lifetime’ of the simulation.

Passengers were accumulated over time using a birth Poisson point process. We then calculated the number of ‘detectable’ passengers that preceded the acquisition of the driver based on whether the underlying clone had expanded to a great enough proportion of HSC cells. We examined the association between the number of detectable passengers and the fitness of the underlying HSC clone. We implemented this simulation in the Julia programming language 1.4(Bezanson *et al.*, 2017).

Data Availability

Individual whole-genome sequence data for TOPMed whole genomes, individual-level harmonized phenotypes and the CHIP variant call sets used in this analysis are available through restricted access via the dbGaP TOPMed Exchange Area available to TOPMed investigators. Controlled-access release to the general scientific community via dbGaP is ongoing. Accession numbers for these datasets are:

Figures



Model fit by covariate sets

Covariates	Rsq(%)	Adjusted-Rsq(%)	AIC
passengers	12.6%	11.0%	139.3
passengers and age*	31.7%	29.1%	127.8
passengers, age*, and VAF*	32.5%	28.6%	129.1

*Recorded from first blood draw

Figure 4.1: Estimating Clonal Expansion using Passenger Counts in TOPMed Genomes.

A, A schematic depiction of using passenger counts to estimate the rate of expansion of a hematopoietic stem cell (HSC) clone that expands due to the acquisition of a driver mutation. The passengers (blue) that precede the driver (red) can be used to date the acquisition of the driver. **B**, The observed clonal expansion rates (dVAFdT), as expressed in the change in variant allele frequency (VAF) over time (years), were associated with increased passenger counts. **C,** A multivariate model including passenger counts, age at blood draw, and VAF indicates the relative contributions of age and VAF over a baseline model. **D,** The relative abundances of passenger counts were estimated for CHIP driver genes with at least 30 cases using a negative binomial regression, adjusting for age at blood draw and study. The coefficients are relative to DNMT3A R882- CHIP.

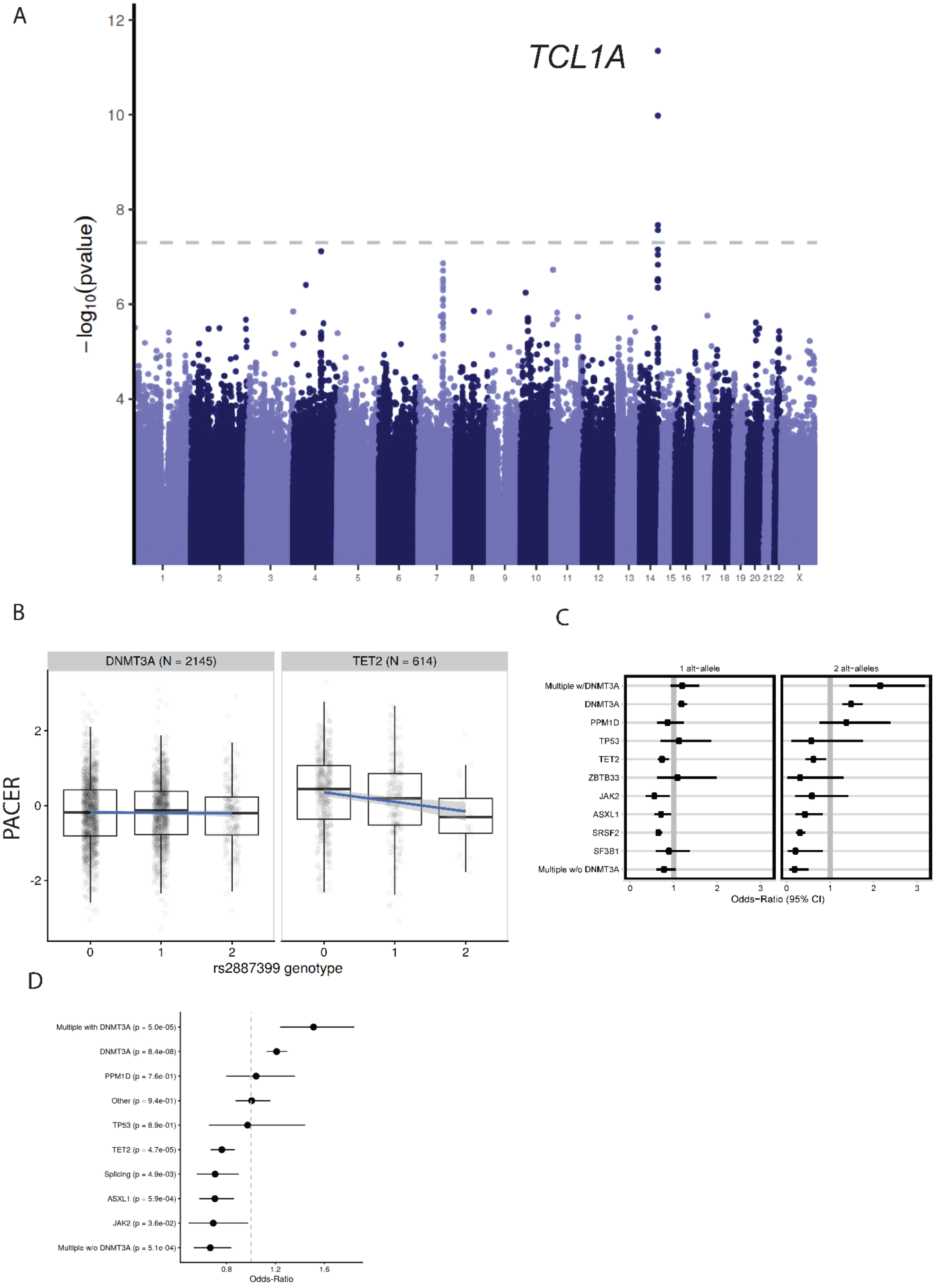

A
12
10
8
6
4
−log10(pvalue)
TCL1A
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
B
DNMT3A (N = 2145)
TET2 (N = 614)
PACER
2
0
-2
0
1
2
0
1
2
rs2887399 genotype
C
1 alt-allele
2 alt-alleles
Multiple w/DNMT3A
DNMT3A
PPM1D
TP53
TET2
ZBTB33
JAK2
ASXL1
SRSF2
SF3B1
Multiple w/o DNMT3A
0
1
2
3
0
1
2
3
Odds−Ratio (95% CI)
D
Multiple with DNMT3A (p = 5.0e-05)
DNMT3A (p = 8.4e-08)
PPM1D (p = 7.6e-01)
Other (p = 9.4e-01)
TP53 (p = 8.9e-01)
TET2 (p = 4.7e-05)
Splicing (p = 4.9e-03)
ASXL1 (p = 5.9e-04)
JAK2 (p = 3.6e-02)
Multiple w/o DNMT3A (p = 5.1e-04)
0.8
1.2
1.6
Odds-Ratio

Figure 4.2: Identifying the Inherited Determinants of Clonal Expansion.

A, A genome-wide association study (GWAS) of passenger counts identifies TCL1A as a genome-wide significant locus. **B,** The association between the genotypes of rs2887399 and passenger counts varied between TET2 and DNMT3A. Alt-alleles were associated with decreased passengers in TET2 mutation carriers, in contrast to DNMT3A carriers, where no association was observed. **C**, The association between alt-alleles at rs2887399 and acquisition of specific CHIP mutations varies by CHIP mutations. Alt-alleles increased risk for acquiring DNMT3A mutations but decreased risk for acquiring splicing mutations. **D**, Previously identified phenotypic associations with rs2887399.

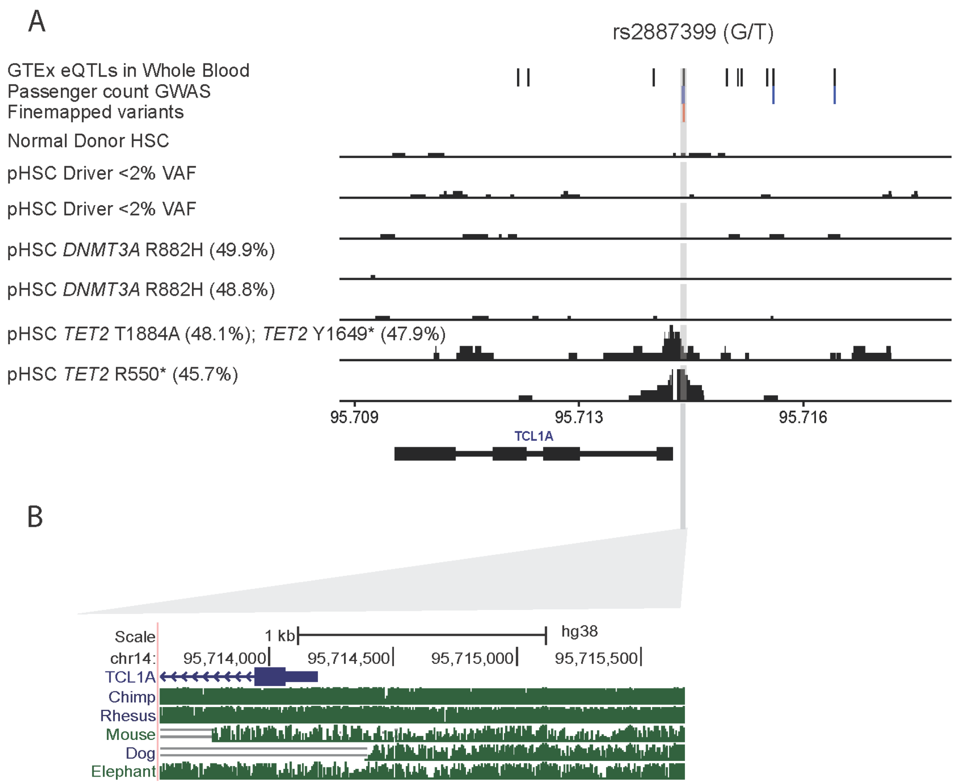
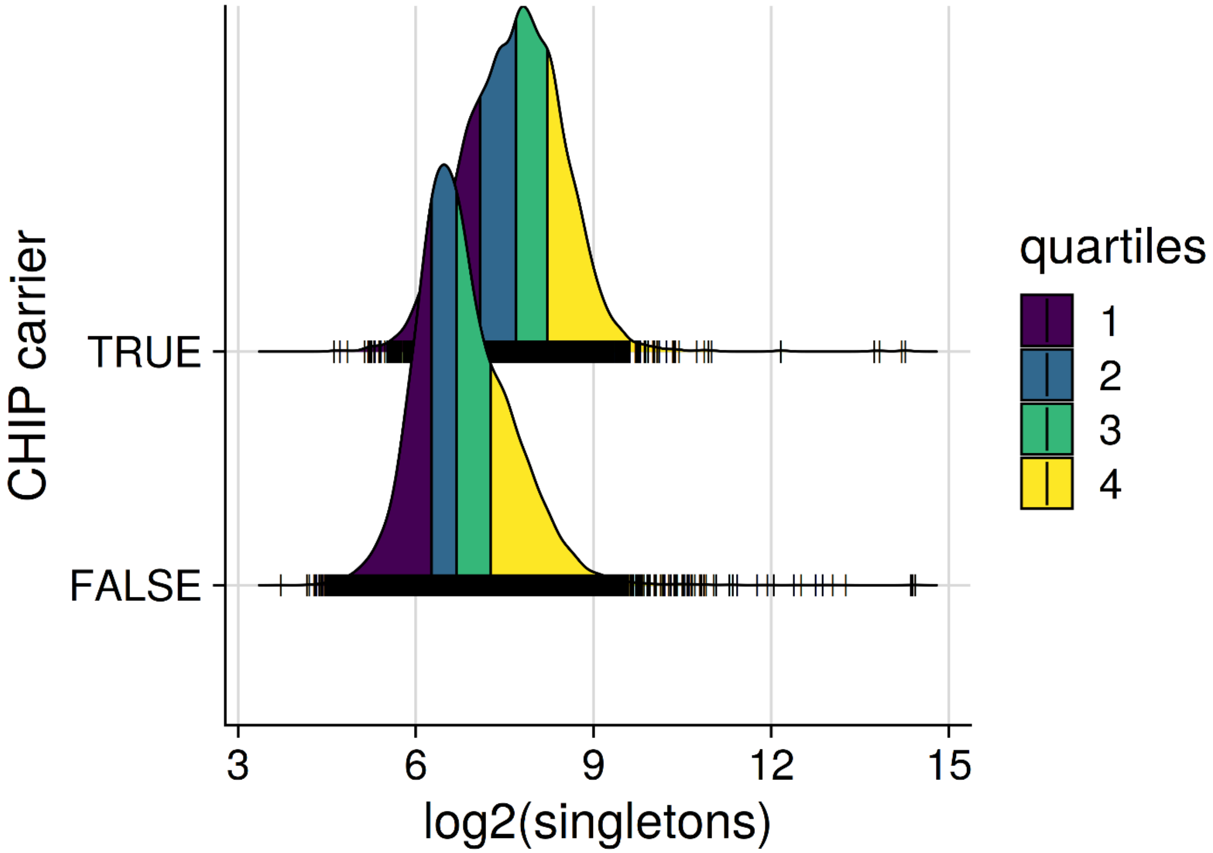


Figure 4.3: Functional Characterization of TCL1A Locus

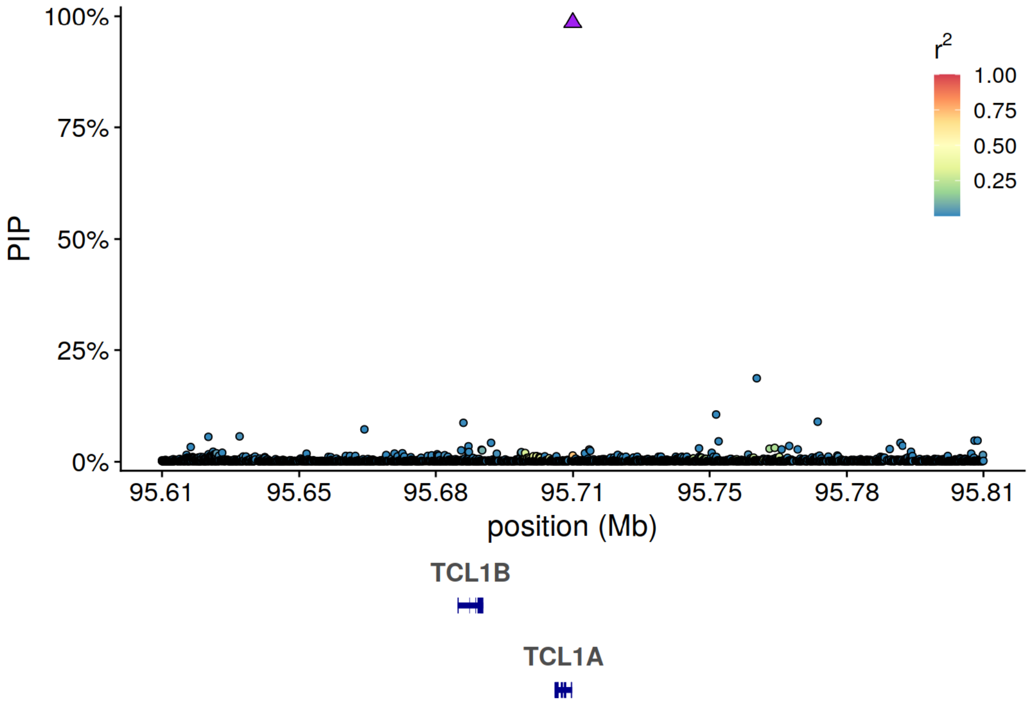
A, Finemapping of the TCL1A locus identified a single causal variant, rs2887399, highlighted with a gray bar. Rs2887399 is an eQTL of TCL1A in whole blood (GTEx v8). The three ATAC-seq panels indicate differential chromatin accessibility at the promoter of TCL1A. **B,** Multiz alignments across multiple species are shown for the TCL1A locus. The TCL1A promoter is not conserved in murine models.

Extended Data Figures



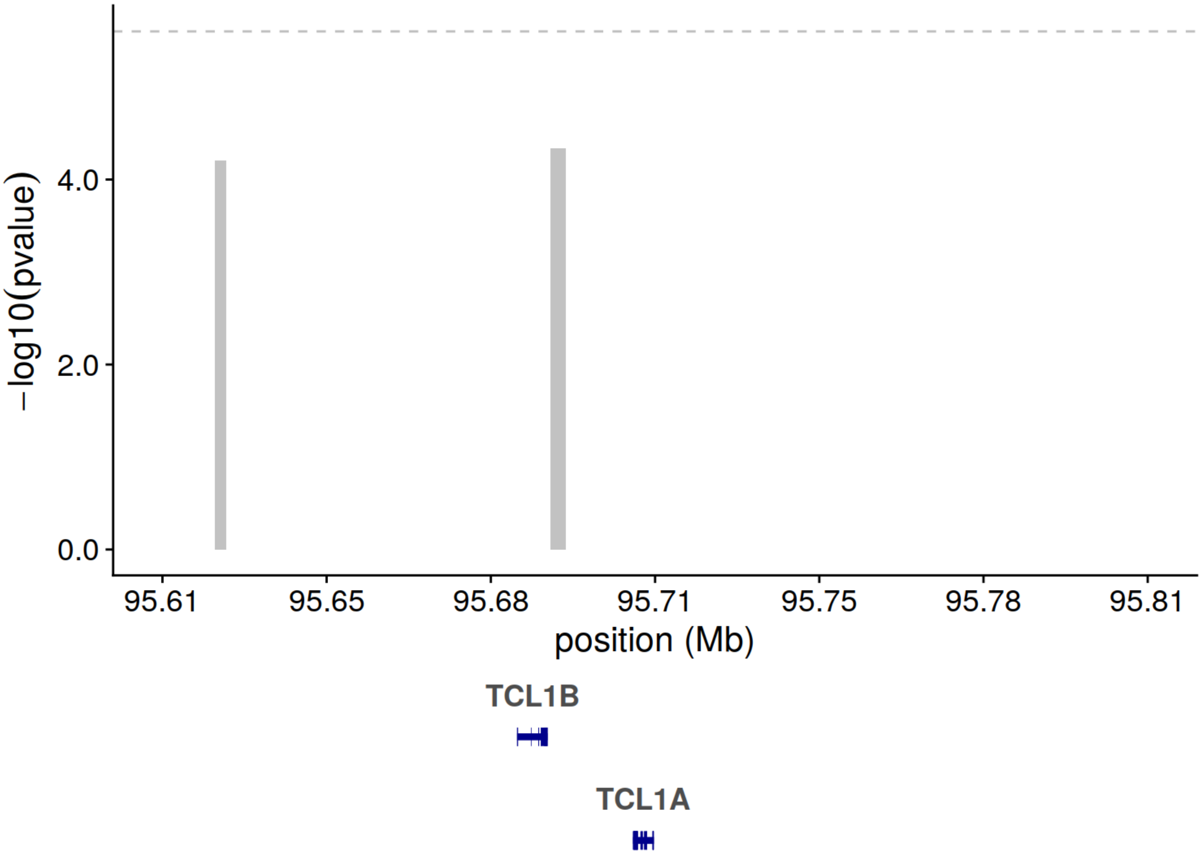
Extended Data Figure 4.4: CHIP carriers are enriched for passengers

The passenger counts are enriched by 54% (95% CI: 51%-57%) after adjusting for age and study using a negative binomial regression. The different colors in the density plots correspond to quartiles of the marginal probability distributions. As the density estimates are smoothed, the underlying data points are indicated with hash marks. The data use a log2 scale, such that an increase by 1 indicates a single doubling has occurred.



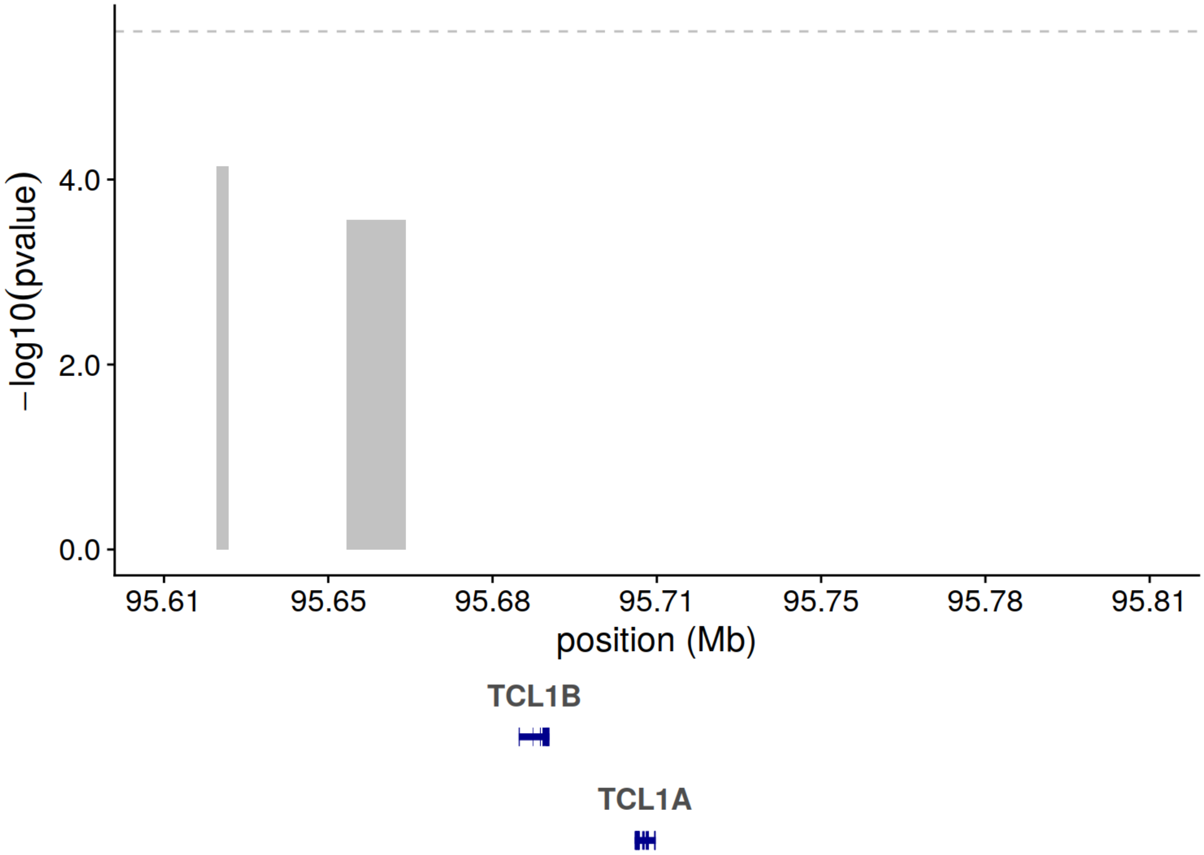
Extended Data Figure 4.5: Finemapping identifies a single causal variant (PIP > 50%) rs2887399.

The posterior inclusion probabilities (PIP) as estimated by SuSIE are plotted on the y-axis, and the genomic position of a 0.8 Mb region including TCL1A is plotted on the x-axis. The linkage disequilibrium (LD) estimates are plotted on a color scale and are estimated on the genotypes used for association analyses.



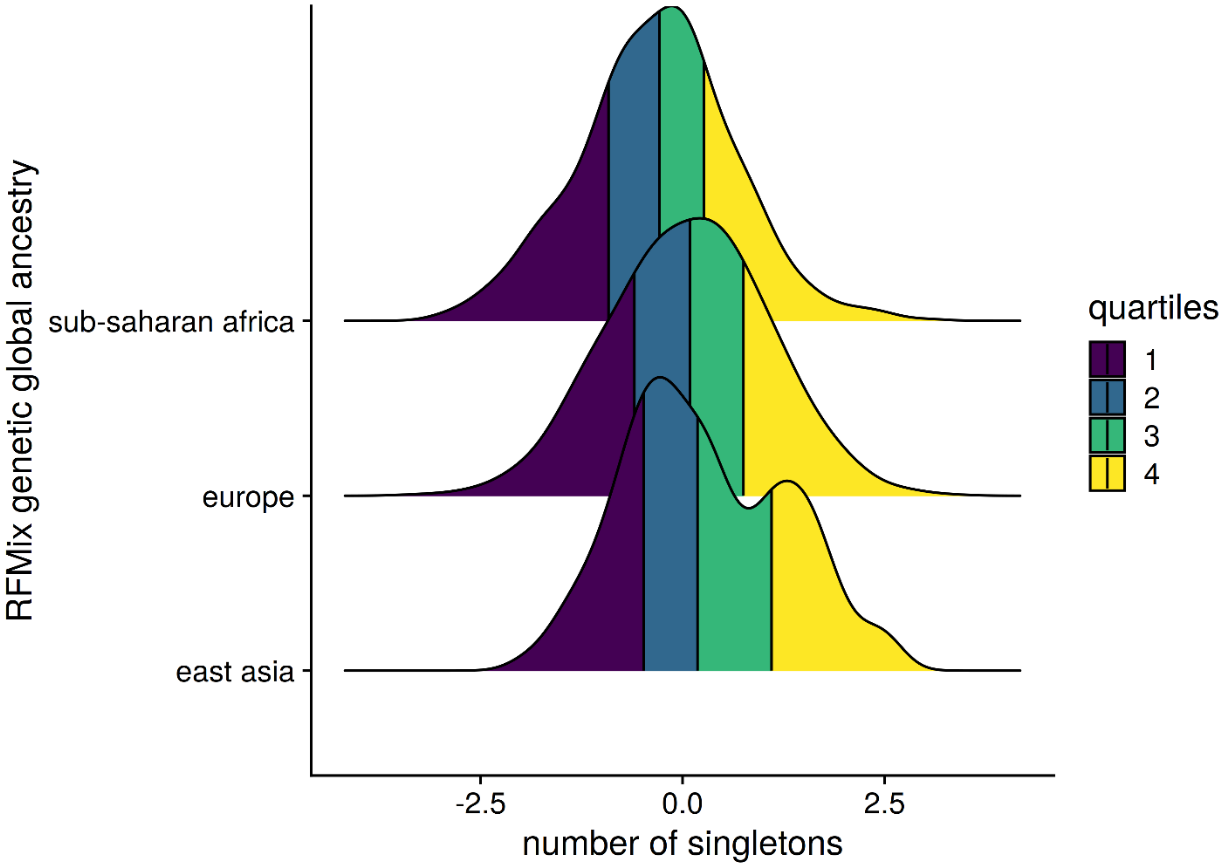
Extended Data Figure 4.6: Rare variant analysis of TCL1A locus identifies a suggestive signal prior to conditioning on rs2887399

Rare variant analyses were performed using the SCANG rare variant scan procedure including all variants with a minor allele count less than 300. Identified rare variant windows are plotted as gray rectangles where the width corresponds to the size of the genomic region and the height corresponds to the pvalue of the SCANG test statistic for the window.



Extended Data Figure 4.7: Conditioning on rs2887399 attenuates independent rare variant signal.

Rare variant analyses were performed including the rs2887399 genotypes as covariate.



Extended Data Figure 4.8: Genetic ancestry is associated with passenger counts

Ancestries were estimated for each genome using RFMix, a supervised machine learning method trained on a reference panel from the Human Genome Diversity Panel. Sub-saharan genomes were modestly depleted of passengers.

Supplementary Information 1: Theoretical Simulation of PACER

PACER Simulation Parameters

We assume that the accumulation of passenger mutations is described by a Poisson birth-death stochastic process. As the birth and death rates scale with the number of HSCs, we assume a linear birth-death process.

We assume that the birth rate for a given hematopoietic stem cell (HSC) i at time t with fitness $s_i(t)$) is $\lambda_i(t) \sim Poisson(\omega * X_i(t) * (1 + s_i(t)) * dt)$, where dt represents the amount of time in years, and ω represents the number of stem cell divisions per year. We assume that the death rate can be described as

$$\psi_i(t) \sim Poisson(\omega * X_i(t) * (1 - s_i(t)) * dt)$$

. The death rate is the rate at which an HSC divides into two differentiated cells, and the birth rate is the rate at which an HSC divides into two HSCs. We don't consider asymmetric HSC differentiation as this would not change the clone size. The HSC clone cell count is defined as $X_i(t) = \sum_{l \leq t} \lambda_i\,(l) - \psi_i(l)$, and the HSC clone size (a fraction of the total cell population) is $VAF_i(t) = \frac{X_i(t)}{\sum_j X_j(t)}$.

We start with 500 HSC clones, each with 200 identical cells in each clone $X_i(t = 0) = 200$. Each cell divides once every three years (= 1/3), and each clone with an initial $s_i(t = 0) = 0$. At each iteration, we also center the $s_i(t)$ such that $\overline{s_i(t)} = 0$. This means that there are 100,000 total HSCs at the start of the simulation.

For each clone, we set the passenger mutation rate:

1. μ_p, the passenger accumulation rate, $A_i(t) \sim Poisson(X_i(t) * \mu_p * dt)$

Where $A_i(t)$ is the number of passengers accumulated in a given clone through time t. We set $\mu_p = 0.006$, which is the passenger mutation rate of a diploid genome for a single HSC per year. This implies a mutation rate of 6 passengers per year for a clone with 1000 cells, and a mutation rate of 600 passengers per year across the entire population of 100,000 HSCs. We will later consider the effects of an insensitive sequencing assay that captures a small fraction of the passengers.

We assign a single driver to one of the HSC clones, which is randomly selected among the HSC clones. The time of acquisition is uniformly drawn from each cell division after 10 years, such that are driver is equally likely to be acquired at either 10 years or 78 years. We simulate the HSC population across a lifetime of 90 years. We refer to the time of driver acquision as T_d.

We assume that each HSC clone can at most acquire a single driver, which represents a similar HSC population to the TOPMed CHIP driver carriers.

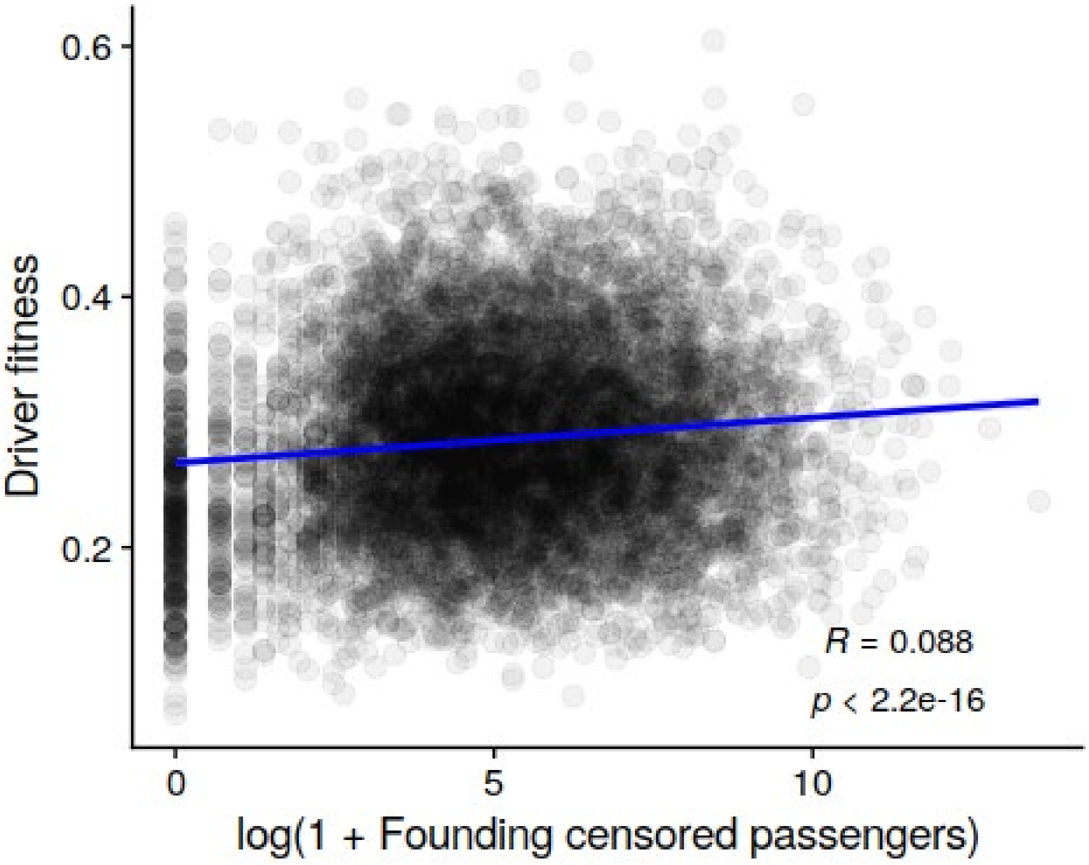
If an HSC clone i acquires a driver at time t, we set $s_i(t) = Beta(4,16)$. A $Beta(4,16)$ random variable is bounded between 0 and 1 and has an expectation of 0.20. An HSC with $_i(t) = 0.20$ will self-renew 60% of the time, and terminally differentiate 40% of the time.

For a given HSC population, we simulate 90 years, and track the accumulation of passengers and drivers. To incorporate the censoring from using 38x sequencing coverage, we simulate whether a given passenger would be observed at 38x coverage by sampling the number of alt-reads from $R \sim Binomial(38, VAF_i(t))$ and comparing $R \geq 2$, since two reads are required by our variant calling process. We refer to $P(R \geq 2 | VAF = vaf) = P(Binomial(38, vaf) \geq 2)$ We refer to the passengers that would be detected at 38x coverage as the censored founding passengers, $AC_i(t)$, where $t = T_d$.

We ran the simulation 10,000 times, where at most a single HSC clone acquires a driver mutation. We then compared $AC_i(T_d)$ to the fitness of the clone at the end of each simulation.

PACER Simulation Results

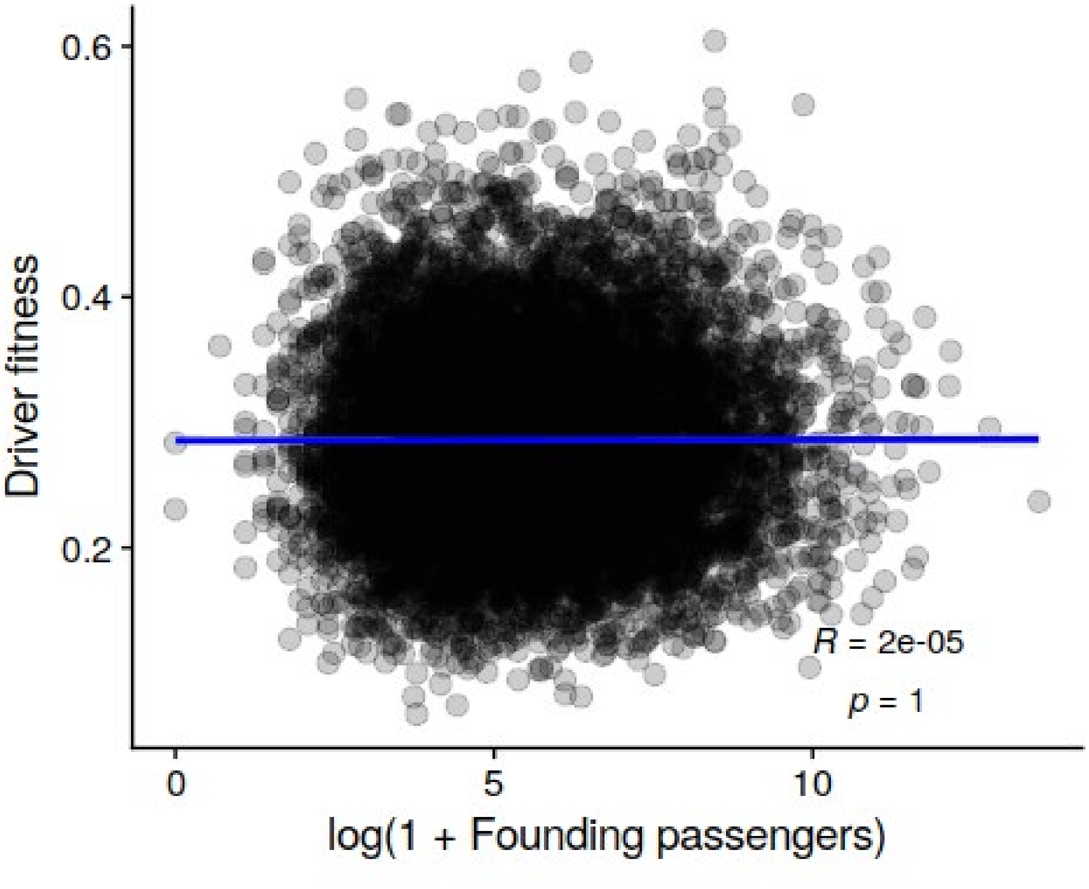
Supplementary Figure S6



We observed a modest concordance between the founding censored passengers with the fitness of each HSC at the end of the simulation (spearman = 0.09, pvalue < 2.2e-16, Supplementary Figure S6). This suggests that the censored founding passengers are proportional to the fitness of the driver mutations. Stochastic drift of the HSC clone sizes contributes substantial residual variance.

Supplementary Figure S7

We observe no concordance between the uncensored founding passengers with fitness (Supplementary Figure S7).



Supplementary Tables

Supplementary Table 4.1: List of CHIP mutations queried

https://drive.google.com/file/d/1WMiMKrRiZjeboRz3Z88lT0LvYN6VryVq/view?usp=sharing

References

Bates,D. *et al.* (2019) Matrix: Sparse and Dense Matrix Classes and Methods.

Bezanson,J. *et al.* (2017) Julia: A fresh approach to numerical computing.

Bick,A.G. *et al.* (2020) Inherited causes of clonal haematopoiesis in 97,691 whole genomes. *Nature*, **586**, 763–768.

Bick Alexander G. *et al.* (2020) Genetic Interleukin 6 Signaling Deficiency Attenuates Cardiovascular Risk in Clonal Hematopoiesis. *Circulation*, **141**, 124–131.

Bycroft,C. *et al.* (2017) Genome-wide genetic data on ~500,000 UK Biobank participants. *bioRxiv*, 166298–166298.

Carvalho-Silva,D. *et al.* (2019) Open Targets Platform: new developments and updates two years on. *Nucleic Acids Res*, **47**, D1056–D1065.

Cibulskis,K. *et al.* (2013) Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology*, **31**, 213–219.

Cingolani,P. *et al.* (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, **6**, 80–92.

Consortium,T.Gte. (2020) The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*, **369**, 1318–1330.

Corces,M.R. *et al.* (2016) Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nature Genetics*, **48**, 1193–1203.

Desai,P. *et al.* (2018) Somatic mutations precede acute myeloid leukemia years before diagnosis. *Nature Medicine*, **24**, 1015–1023.

Dr,Z. *et al.* (2015) The ensembl regulatory build. *Genome Biol*, **16**, 56–56.

Fishilevich,S. *et al.* (2017) GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database (Oxford)*, **2017**.

Genovese,G. *et al.* (2014) Clonal Hematopoiesis and Blood-Cancer Risk Inferred from Blood DNA Sequence. *New England Journal of Medicine*, **371**, 2477–2487.

Gogarten,S.M. *et al.* (2019) Genetic association testing using the GENESIS R/Bioconductor package. *Bioinformatics*, **35**, 5346–5348.

Hecht,F. *et al.* (1984) Common region on chromosome 14 in T-cell leukemia and lymphoma. *Science*, **226**, 1445–1447.

Hiatt,J.B. *et al.* (2013) Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation. *Genome Res.*, **23**, 843–854.

Jaiswal,S. *et al.* (2014) Age-Related Clonal Hematopoiesis Associated with Adverse Outcomes A BS TR AC T. *NEJM.org. N Engl J Med*, **26**, 2488–98.

Jaiswal,S. *et al.* (2017) Clonal Hematopoiesis and Risk of Atherosclerotic Cardiovascular Disease. *New England Journal of Medicine*.

Jun,G. *et al.* (2015) An efficient and scalable analysis framework for variant extraction and refinement from population scale DNA sequence data. *Genome Res.*, gr.176552.114.

Koboldt,D.C. *et al.* (2012) VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.*, **22**, 568–576.

Laine,J. *et al.* (2000) The Protooncogene TCL1 Is an Akt Kinase Coactivator. *Molecular Cell*, **6**, 395–407.

Lee-Six,H. *et al.* (2018) Population dynamics of normal human blood inferred from somatic mutations. *Nature*, **561**, 473–478.

Li,Z. *et al.* (2019) Dynamic Scan Procedure for Detecting Rare-Variant Association Regions in Whole-Genome Sequencing Studies. *The American Journal of Human Genetics*, **104**, 802–814.

Ma,C. *et al.* (2013) Recommended Joint and Meta-Analysis Strategies for Case-Control Association Testing of Single Low-Count Variants. *Genetic Epidemiology*, **37**, 539–550.

Maples,B.K. *et al.* (2013) RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference. *The American Journal of Human Genetics*, **93**, 278–288.

Osorio,F.G. *et al.* (2018) Somatic Mutations Reveal Lineage Relationships and Age-Related Mutagenesis in Human Hematopoiesis. *Cell Reports*, **25**, 2308-2316.e4.

Pedersen,B.S. and Quinlan,A.R. (2017) cyvcf2: fast, flexible variant analysis with Python. *Bioinformatics*, **33**, 1867–1869.

R Core Team (2020) R: A Language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Regier,A.A. *et al.* (2018) Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. *Nature Communications*, **9**, 1–8.

Robinson,J.T. *et al.* (2011) Integrative genomics viewer. *Nature Biotechnology*, **29**, 24–26.

Stan Development Team (2020a) RStan: The R interface to Stan.

Stan Development Team (2020b) Stan Modeling Language Users Guide and Reference Manual, 2.17.

Steensma,D.P. *et al.* (2015) Clonal hematopoiesis of indeterminate potential and its distinction from myelodysplastic syndromes. *Blood*, **126**, 9–16.

Taliun,D. *et al.* (2019) Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program Genomics.

Thompson,D.J. *et al.* (2019) Genetic predisposition to mosaic Y chromosome loss in blood. *Nature*, **575**, 652–657.

VariantKey: A Reversible Numerical Representation of Human Genetic Variants | bioRxiv.

Venables,W.N. and Ripley,B.D. (2002) Modern Applied Statistics with S 4th ed. Springer-Verlag, New York.

Voss,K. *et al.* (2017) Full-stack genomics pipelining with GATK4 + WDL + Cromwell. F1000 Research.

Wang,G. *et al.* A simple new approach to variable selection in regression, with application to genetic fine-mapping *.

Watson,C.J. *et al.* (2020) The evolutionary dynamics and fitness landscape of clonal hematopoiesis. *Science*, **367**, 1449–1454.

Xie,M. *et al.* (2014) Age-related mutations associated with clonal hematopoietic expansion and malignancies. *Nature Medicine*, **20**, 1472–1478.

Zeller,C. *et al.* (2003) SASH1 : a candidate tumor suppressor gene on chromosome 6q24.3 is downregulated in breast cancer. *Oncogene*, **22**, 2972–2983.

Zhao,Z. *et al.* (2020) UK Biobank Whole-Exome Sequence Binary Phenome Analysis with Robust Region-Based Rare-Variant Test. *The American Journal of Human Genetics*, **106**, 3–12.

Zhou,W. *et al.* (2018) Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature Genetics*, **50**, 1335–1341.

Zhou,W. *et al.* (2016) Mosaic loss of chromosome Y is associated with common variation near TCL1A. *Nature Genetics*, **48**, 563–568.

Chapter 5 Conclusion

In this dissertation, we propose three contributions to modern statistical challenges in human genetics. These contributions map onto two principal axes – 1) how do we leverage EHR-linked biobanks to advance association discovery and characterize disease etiology 2) how can we use population scale whole genome sequencing efforts to characterize the dynamic genome of blood cells. Towards our first axis, we propose study design recommendations for conducting GWAS of EHR derived lab traits and develop a novel MR method for inferring the causal relationship between two traits. Towards our second, we leverage over 100,000 whole genomes from the NHLBI Trans-Omics for Precision Medicine program (TOPMed) (Taliun *et al.*, 2021) to identify ~5,000 carriers of specific class of oncogenic mosaicism in blood and develop a Bayesian model of clonal expansion. Here, we briefly review these contributions, and embed them with the broader landscape of biostatistics and human genetics. We discuss limitations and potential future extensions.

LabWAS and study design recommendations

In chapter 2, we introduce the notion of LabWAS, which is a PheWAS (Denny *et al.*, 2010) analysis of clinical lab traits. To our knowledge, this study is the first to comprehensively conduct parallel analyses of study design decisions for GWAS of EHR derived lab traits. This study utilized GWAS from 70 lab traits from two independent biobanks – the Michigan Genomics Initiative, and BioVU from the Vanderbilt health system. The work benefited from an interdisciplinary team of statistical genetics, genetic epidemiologists, and clinical pathologists.

Among the impetus for this work was the dissensus regarding basic analytic decisions that we observed in the literature. Previous efforts varied greatly in choice of summary statistics and covariate adjustment. When these decisions were emphasized, justifications were often tailored to a single biobank. Permitted by the fortuitous availability of two independent biobanks, we were able to systematically assess the concordance of different analytic decisions. We concluded with a set of recommendations that may be generally appropriate for lab trait GWAS but emphasize that there is no single optimal analysis pipeline. There was considerable heterogeneity in the effects of analysis decisions across the labs. For example, the maximum measurement increased empirical power relative to using the median for a subset of blood cell indices, but decreased power elsewhere.

The future bears no shortage of additional biobanking efforts. The All of Us Initiative from the NIH (The "All of Us" Research Program, 2019) aims to sequence ~1,000,000 genomes and link them to phenotypes. As these efforts accelerate, meta-analyses between biobanks will become more common place. We anticipate that our results and study design recommendations will benefit these future endeavors.

We briefly remark that although the conclusions are derived from two biobanks, the ability to generalize our recommendations to other biobanks remains unclear. Future work is needed to expand the study design analyses across more than two biobanks. Such an expanded analysis may further identify which analytic strategies transcend biobank differences and which should be tailored to individual health systems.

Sparse Mendelian Randomization

Permitted by the accessibility of GWAS summary statistics, analyses that infer causal effects between two traits have become commonplace. These methods incorporate assumptions regarding the genetic architecture of analyzed traits. As traits vary in genetic architecture, bespoke MR methods can be optimized for application. In Chapter 3, we introduce a sparse MR method called SPARMR (SPARse Mendelian Randomization), for application to traits where a dense genetic architecture is ill-suited. In concert with recent advances in fine-mapping methods (Wang *et al.*, 2020), we develop a method that assumes most variants are not causal. We present simulations that demonstrate the performance of the method in various settings.

The method presented here is enabled by the advent of probabilistic programming languages (PPL), which make random variables first class citizens. Here, we implemented our method using Tensorflow probability (TFP) (Dillon *et al.*, 2017), which abstracts away model definition from MCMC sampling. TFP facilitates efficient sampling using Hamiltonian Monte-Carlo, which typically out performs traditional MCMC approaches in many common settings (Betancourt, 2018). PPLs have become increasingly common in genomics (Berzuini *et al.*, 2020), as their flexibility and performance facilitates the rapid development of bespoke statistical models.

We remark that despite recent advances in MR methods, causal inference between two traits using GWAS summary statistics remains challenging. Disease associations are often highly pleiotropic (Morrison *et al.*, 2020; Verbanck *et al.*, 2018), and incomplete knowledge of all relevant confounders makes causal interpretation challenging unless the disease system is very well characterized. Although methods that make inference more robust to pleiotropy mitigate this to some degree, we submit that study design has a much larger role in the application of MR.

Using WGS to study oncogenic mosaicism in blood

In chapter 4, we called somatic mutations genome wide across ~127,000 samples from the Trans-Omics for Precision Medicine (TOPMed) consortium (Taliun *et al.*, 2021). These samples were extracted primarily from whole blood and were sequenced to 38x coverage on average. Cross-referencing these somatic variant calls identified ~5,000 carriers of a leukemogenic class of point mutations – a state called clonal hematopoiesis of indeterminate potential (CHIP). The landscape of somatic variation is diverse, and includes several variants with no fitness effects, which are sometimes referred to as passengers. Here we used the passengers to construct a mutational clock. As the data are likely generated by an underlying stochastic mutational process, we developed a Bayesian hierarchical model to describe the underlying process for passenger mutation acquisition.

Our estimates are limited by the sensitivity of 38x WGS, which was not collected with somatic variant calling in mind. We called variants using only a single tissue sample, which likely enriches our call-set for artifacts relative to traditional paired-tumor somatic variant calling. We anticipate that as paired-tumor samples become increasingly available, similar analyses may become possible with a refined somatic variant call-set.

To date, no therapeutic interventions exist for CHIP, despite its phenotypic consequences. We anticipate that as molecular pathways that modulate clonal expansion are identified, promising new avenues for diagnostics and therapeutics will arise. We speculate that detection of oncogenic mosaicism in the blood of older individuals may someday become routine in clinical care.

Summary

The current era represents an exciting opportunity for computational genetics. As sequencing costs plummet, a future where human genomes are routinely sequenced seems eminently possible. In this dissertation, I make three contributions towards addressing the challenges that result from this data deluge. With eagerness and curiosity, I look forward to future progress.

References

Berzuini,C. *et al.* (2020) A Bayesian approach to Mendelian randomization with multiple pleiotropic variants. *Biostatistics*, **21**, 86–101.

Betancourt,M. (2018) A Conceptual Introduction to Hamiltonian Monte Carlo. *arXiv:1701.02434 [stat]*.

Denny,J.C. *et al.* (2010) PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics*, **26**, 1205–1210.

Dillon,J.V. *et al.* (2017) TensorFlow Distributions. *arXiv:1711.10604 [cs, stat]*.

Morrison,J. *et al.* (2020) Mendelian randomization accounting for correlated and uncorrelated pleiotropic effects using genome-wide summary statistics. *Nature Genetics*, **52**, 740–747.

Taliun,D. *et al.* (2021) Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature*, **590**, 290–299.

The “All of Us” Research Program (2019) *New England Journal of Medicine*, **381**, 668–676.

Verbanck,M. *et al.* (2018) Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nature Genetics*, **50**, 693–698.

Wang,G. *et al.* (2020) A simple new approach to variable selection in regression, with application to genetic fine mapping. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **82**, 1273–1300.