Title
Understanding the genetic architecture of complex traits through meta-analysis

Author
Taraszka, Kodi

Publication Date
2022

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Understanding the genetic architecture of complex traits through meta-analysis

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Computer Science

by

Kodi Nicole Taraszka

2022

ABSTRACT OF THE DISSERTATION

Understanding the genetic architecture
of complex traits through meta-analysis

by

Kodi Nicole Taraszka
Doctor of Philosophy in Computer Science
University of California, Los Angeles, 2022
Professor Eleazar Eskin, Chair

Exploring how genetic architecture shapes complex traits and diseases is a central premise of human genetics. Over the years, genome-wide association studies (GWAS) have enabled the discovery of numerous genetic variants associated with a variety of complex traits. In addition to the large array of traits analyzed, GWAS in diverse ancestral populations have also seen a significant increase in sample sizes. These efforts led to tens of thousands of publicly available GWAS summary statistics whose known correlation structure could be leveraged for further discovery.

In this dissertation, I present two novel methods for the meta-analysis of GWAS summary statistics as well as conduct a pan-cancer meta-analysis of somatic variant burden. For one method, I present a likelihood ratio test for the joint analysis of genetically correlated traits and provide a per trait interpretation framework of the omnibus association. For the other method, I present a Bayesian framework that improves fine mapping of significant associations for one trait by leveraging the complementary information from distinct ancestral backgrounds. In addition to these methods, I analyzed how clinical and polygenic germline features influence somatic variant burden within and across cancer types.

The dissertation of Kodi Nicole Taraszka is approved.

Noah A. Zaitlen

Jason Ernst

Sriram Sankararaman

Eleazar Eskin, Committee Chair

University of California, Los Angeles

2022

To my family

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First, I would like to thank my advisor, Professor Eleazar Eskin. Eleazar was my primary advisor at the beginning of my PhD, and it was under his guidance that I learned about the field of quantitative genetics. He helped me understand the value of beginning with first principles and building up in complexity. He also helped sharpen many of my soft skills. I deeply appreciate how much I grew during my time in Zarlab.

I would also like to thank my co-advisor, Professor Noah Zaitlen, who was my primary advisor during the latter half of my PhD. It was under his guidance that I grew into an independent scientist. Noah helped me further hone many of my soft skills particularly, how I give presentations and write papers. I would also like to acknowledge my other committee members, Professor Sriram Sankararaman and Professor Jason Ernst for their support and guidance.

Professor Alexander "Sasha" Gusev was an informal co-advisor during the last three years of my PhD. Through my collaboration with Sasha Gusev and Noah Zaitlen, I found a number of interesting questions that I am deeply passionate about. Sasha has had a great deal of impact on my scientific interests and my professional career for which I am incredibly grateful. I would also like to thank Professor Tandy Warnow, my undergraduate research advisor at the University of Illinois Urbana-Champaign. It was through working with Tandy that I first discovered my interest in Bioinformatics and methods development. I am beyond grateful I had the privilege of working with Tandy and appreciate her continued mentorship.

I would like to also thank my collaborators, including Professor Eleazar Eskin, Professor Noah Zaitlen, Nathan LaPierre, Helen Huang, Rosemary He, Farhad Hormozdiari, Professor Serghei Mangul, Professor Alexander Gusev, Stefan Groha, Yixuan He, and Kathleen Houlahan. Additionally, this endeavor would not have been possible without my collaboration with Tempus Labs, Inc., particularly David King, Robert Tell, and Kevin White. I am so grateful for their contributions which were critical for my success.

Overall, I am truly grateful that I was a member of two positive and supportive labs. My

lab mates were some of my favorite people at UCLA. I would like to thank Nathan LaPierre, Laura Kim, Rosemary He, Varuni Sarwal, Dat Duong, Lisa Gai, Jennifer Zou, Serghei Mangul, Robert Brown, Jeremy Rotman, Ariel Wu, Christa Caggiano, Ella Petter, Joel Mefford, Michal Sadowski, Jamie Matthews, Michael Thompson, Richard Border, Angela Wei, Ryo Yamamoto, and Terence Li. I would also like to thank the Bioinformatics community at UCLA, particularly my year's cohort, for adopting the Computer Science PhD students into their cohort.

Finally, and most importantly, I would like to thank my family. I would like to thank my husband, Paul, who has been a pillar of support through everything. I would also like to thank my daughter, Eleanor, who has brought such joy to my life. I would also like to thank my parents and in-laws: Kirk Collins, Rhonda and Brandon Thomas, and Jan and Lucyna Taraszka, for their love and support. I would also like to thank my siblings as well as acknowledge my grandparents and extended family.

VITA

2012 – 2017	B.S., Statistics and Economics, University of Illinois, Urbana, Illinois, USA.
2019 – 2019	Visiting Graduate Researcher, Dana Farber Cancer Institute, Boston, Massachusetts, USA
2020 – 2020	Visiting Intern, Tempus Labs, Inc., Chicago, Illinois, USA.
2018 – 2019	Graduate Teaching Assistant, University of California Los Angeles, California, USA
2017 – 2022	Graduate Research Assistant, University of California, Los Angeles, California, USA

PUBLICATIONS

K. Taraszka, S. Groha, D. King, R. Tell, *et al.* Pan-cancer analysis of clinical and genetic determinants of somatic mutational burden. *in preparation.*

Y. He*, S. Groha*, **K. Taraszka***, C.M. Lakhani, *et al.* Genetic Ancestry and Population Differences in Somatic Alterations and Clinical Outcomes for Five Common Cancers. *in preparation.*(* Authors contributed equally)

K.E. Houlahan, J. Yuan, T. Schwarz, J. Livingstone, *et al* [including **K. Taraszka**]. Germline determinants of the prostate tumor genome. *in preparation.*

K. Taraszka, N. Zaitlen, E. Eskin. Leveraging pleiotropy in genome-wide association studies

in multiple traits with per trait interpretations. *PLOS Genet* **18**, 11 (2022).

A.H. Nassar*, E. Adib*, S. Abou Alaiwi*, T. El Zarif, *et al.* [including **K. Taraszka**]. Integrated Clinico-genomic Analysis of Genetic Ancestry in Patients with Solid tumors treated with Immune Checkpoint Inhibitors. *Cancer Cell* **40**, 10 (2022). (* Authors contributed equally).

L. Gai*, J. Karlin*, N. LaPierre, K. Danesh, *et al* [including **K. Taraszka**]. Deep neural network guided detection of thyroid eye disease from external photos. *British Journal of Ophthalmology in press.* (* Authors contributed equally)

A. Gusev, S. Groha, **K. Taraszka**, Y.R. Semenov, *et al.* Constructing germline research cohorts from the discarded reads of clinical tumor sequences. *Genome Med* **13**, 179(2021).

N. LaPierre*, **K. Taraszka***, H. Huang, R. He , *et al.* Identifying causal variants by fine mapping across multiple studies. *PLOS Genet* **17**, 9 (2021). (* Authors contributed equally)

M. Alser*, J. Rotman*, D. Deshpande, **K. Taraszka**, *et al.* Technology dictates algorithms: recent developments in read alignment. *Genome Biol* **22**, 249 (2021). (* Authors contributed equally)

K. Collins, T. Warnow. Pasta for proteins. *Bioinformatics* **34**, 22 (2018).

CHAPTER 1

Introduction

Scientists have long noted that measurable traits, such as height and disease risk, are correlated within families. They have also linked many physical and psychological phenotypes to genes; however, it was only after the invention of high throughput sequencing was it feasible to tie genetic variants and complex traits at scale. In part, this is because the human genome consists of approximately three billion nucleotides composed of four types of bases: Adenine (A), Cytosine (C), Guanine (G), and Thymine (T). DNA is physically organized as a double helix structure connected by the base pairs, where A bonds with T and C bonds with G. In humans, the genome is split and compressed into twenty three segments called chromosomes. As humans are a diploid species, every somatic cell (i.e. non egg or sperm cell) contains two complete copies of each chromosome, with each parent contributing one copy. Within each chromosome nucleotides are organized into genes, which are then transcribed into RNA which is then translated into proteins; these proteins are responsible for every function in the cell.

While there is significant phenotypic variability amongst humans, there is very little genetic variation within between genomes relative to its size. In fact, humans are almost genetically identical with only 0.1% of DNA differing between individuals. These genetic differences are introduced through mutational events such as single nucleotide polymorphisms (SNPs) and copy number variants (CNVs). A SNP is the most common type of genetic variant and represents the substitution of a single base pair (e.g. A changing to C). CNVs are also common but are a more complex type of alteration which represents the deletion or duplication of genomic regions spanning fifty or more base pairs. Both of these genetic alterations as well as many others are present in all cells including gametes (i.e. egg and

sperm cells) meaning they will be inherited by offspring. In addition to heritable genetic alterations which are often referred to as germline variants, mutations may also be introduced within an individual's lifetime. These alterations are called somatic mutations and may also take a number of forms, including single nucleotide variants and copy number alterations. Somatic variants are induced by a number of causes including exogenous factors, such as tobacco smoke and ultraviolet radiation, as well as through endogenous defects in DNA mismatch repair and DNA replication. While these mutations will be present in the daughter cells of the originating cell, they will not be passed along to any offspring.

In order to tie genetic variants to health outcomes, scientists introduced genome-wide association studies (GWAS); a modern approach for understanding the impact of germline variants on complex traits [157]. To make the large sample sizes required for GWAS at minimal costs, scientists genotype a subset of SNPs spread across the genome and then impute the remaining SNPs using the correlation structure between SNPs known as linkage disequilibrium (LD). As neighboring SNPs tend to be inherited together, their variants or alleles are highly correlated. Every SNP is then converted to a numerical value by counting how many copies each person has of the less frequent base pair known as the minor allele (i.e. 0, 1 or 2 copies). This count is then scaled and centered according to the minor allele frequency. Each SNP is then independently tested using a linear model to determine if it is correlated with the quantitative phenotype. In addition to testing the additive effect of individual variants, polygenic risk scores (PRS) are used to estimate the collective effect of genetic variants genome-wide. These linear predictors can then be used to determine relative risk within a cohort or population.

While a significant amount of research has focused on the role germline variants have regarding health outcomes, scientists have also investigated how somatic mutations impact disease risk and prognosis. While some somatic mutations may be "hotspot" mutations indicating that a particular variant is present in many individuals, most somatic alterations are rare events and likely to only be present in a few individuals. In fact, they may even only be present in a subset of the cells collected for sequencing. As a result, somatic variants are

identified through whole exome sequencing (WES) which targets the exome of all genes or through a panel targeting a particular subset of genes [10, 52]. Similar to the processing of germline variants, somatic alterations are converted to a numerical value, typically a binary variable indicating their presence. They are then tested using a linear model for association with the trait of interest.

Overall, technological advancements have resulted in a continual decrease in sequencing costs enabling a steady increase in both the number of studies and their sample sizes. This has led to multiple large biobanks through which tens of thousands of unique germline variants have been associated with a variety of phenotypes [147]. In addition to the large array of traits analyzed, GWAS have also expanded to better represent diverse ancestral populations. These efforts have resulted in a plethora of publicly available GWAS summary statistics measuring hundreds of thousands of individuals from around the globe [131, 147]. The same explosion in data has also taken hold in cohorts of somatic sequencing. In particular, a recent imputation procedure was introduced that generates cohorts with germline and somatic calling but only requires the direct sequencing of somatic variants [60].

While the initial analyses in these cohorts have led to a better understanding of the additive effect of genetic variants on complex traits and diseases, the complete genetic architecture still remains unclear. In part, this is because much of the research effort has focused on common variants or SNPs present in at least 1% of individuals; however, research has shown that rare variants have an appreciable effect on common phenotypes and ailments. Another contributing factor is that many traits are polygenic which means that multiple variants have a small influence on the trait [157]. Unfortunately, without sufficiently large sample sizes, these weak effects cannot be discovered [113].

While scientists have been able to characterize the relationship between genetic variation and complex traits by increasing both the sample sizes and the number of traits analyzed in studies, there still remains untapped knowledge within the currently available data that could be uncovered via methodological advancements. For example, GWAS discovery power has increased over time resulting in more variants with small effect sizes being discovered.

This suggests the presence of additional small effect variants that have not been identified due to statistical power. Additionally, many variants affect more than one trait, a phenomenon known as pleiotropy) [27, 56, 144, 156, 160]. By modeling this shared genetic architecture, computational methods could have increased power to discover genetically correlated variants.

Another correlation structure that could be modeled by computational methods is the LD between SNPs [157]. Currently, LD results in non-causal SNPs to be reported as associated with the trait because they are merely correlated with the causal SNP(s). These spurious associations make it difficult to prioritize SNPs for experimental follow-up, due to the search space of associated SNPs simply being too large and noisy. While a number of methods have been introduced to reduce the set of associated SNPs using the procedure known as fine-mapping, the set of prioritized SNPs still contain many spurious associations [13, 68, 76, 97, 109]. Furthermore, GWAS have expanded to study a variety of ancestral populations and LD patterns vary across ancestry; methods can utilize this information to better refine the set of candidate SNPs.

Lastly, previous methodological work has generated large cohorts with both germline and somatic variants called [60]. These cohorts can be used not only for further methodological advancements but to directly study the relationship between germline and somatic variants. By conducting novel analyses such as GWAS and PRS, scientists can begin to uncover how germline variants influence somatic alterations. Additionally, both germline and somatic alterations have separately been linked to diseases such as cancer; however, there remains an open question of whether these genetic features jointly influence outcomes or even modify the other's impact [24, 40, 47, 58, 79, 122, 123]. Furthermore, their influence on prognosis both separately and jointly is also an understudied subject.

In this dissertation, I introduce two statistical methods as well as conduct a novel discovery project. In chapter 2, I introduce a method for meta-analyzing GWAS summary statistics. My method, PAT, leverages the pleiotropy between traits to improve statistical power and conducts an omnibus association test. In addition to introducing a novel method,

I also extend the previously developed m-value framework to a multi-trait model. By computing a m-value for all significantly associated variants, I provide a per trait interpretation of the signal. In chapter 3, I introduce MsCAVIAR, a Bayesian method for fine-mapping the set of associated SNPs. Our method meta-analyzes multiple ancestral populations to generate a subset of the associated variants most likely to be causal, known as the causal set. MsCAVIAR extends the CAVIAR framework by modeling the LD across cohorts which enables a further reduction of the causal set size. Lastly in chapter 4, I analyze how clinical and germline features influence the accumulation of somatic variants within individual cancers and across cancer types via a meta-analysis. I established that clinical features, fine scale genetic ancestry and polygenic risk scores shape the somatic landscape both pan-cancer and within specific cancers and find that these associations have implications for survival. Overall, my work presented in the following chapters advances the research communities' understanding of the genetic architecture of complex traits and diseases through meta-analysis.

CHAPTER 2

Leveraging pleiotropy for joint analysis of genome-wide association studies with per trait interpretations

2.1 Introduction

Genome-wide association studies (GWAS) have been instrumental in identifying genetic variants associated with complex traits [39, 42, 104]. As a result, there are tens of thousands of unique associations in the GWAS catalog [96]. With ever increasing sample sizes in GWAS, more and more associated variants have been discovered. This suggests the presence of a large number of variants with small effect sizes that are not identified due to statistical power [113]. With the number of traits examined as well as sample sizes increasing over time, numerous variants are observed affecting more than one trait (i.e., pleiotropy) [27, 56, 144, 156, 160]. Some examples of pleiotropic effects include muscle mass and bone geometry, male pattern baldness and bone mineral density, as well as between multiple psychiatric disorders [32, 74, 169].

We hypothesize that because variants often affect more than one trait, we can leverage this pleiotropy to jointly analyze multiple traits. This would potentially increase statistical power and identify variants with even weaker effect sizes. Following this intuition, there have been many approaches for performing association tests using summary statistics across multiple traits [14, 15, 49, 50, 51, 84, 91, 115, 126, 151, 172, 174, 175]. While simultaneously analyzing multiple traits is advantageous for identifying novel variants, performing an omnibus test is inherently difficult to interpret. This is because an omnibus test assigns one p-value per variant for the set of traits, and it is not clear how to assign a per trait significance level

in this context. Even when this is done, it is not straightforward to interpret due to issues such as inflation in false discovery rates when the assumption of homogeneity in effect sizes is violated [151].

In this chapter, we propose an alternative framework with a two step procedure. First, all traits are jointly analyzed to produce one p-value for each variant. If this p-value is significant, it suggests that the variant is associated with one or more of the traits. To accomplish this first step, we develop an efficient method called pleiotropic association test (PAT) which leverages the estimated genome-wide genetic correlation between traits to improve power and uses null simulations to accurately calibrate p-values. PAT also utilizes importance sampling to allow for estimation of significant p-values efficiently. The second step builds upon an interpretation framework first developed in the context of meta-analysis, m-values, to compute the posterior probability that a variant is associated with each trait [63]. We extend the m-value framework to take into account environmental and genetic correlation between traits.

In simulated data reflecting estimates of genetic and environmental covariance between real UK Biobank traits, we find that PAT is able to correctly control for false positives and increase power to identify novel associations [110, 111, 147]. In comparisons to three multi-trait methods, MTAG, HIPO and ASSET, PAT has a 15.3% increase in the number of associations over the next best method [14, 126, 151]. These results were then interpreted using the m-value framework where PAT identified 37.5% more per trait associations. Additionally while HIPO has only a 16.0% increase in power relative to MTAG for omnibus association testing, using the m-value framework to interpret HIPO's associations resulted in a 46.6% increase in per trait associations relative to MTAG. Finally, we analyzed four traits in the UK Biobank where PAT identified 22,095 novel variants and interpret the results for every trait using m-values. In two of the four traits, the number of per trait associations was almost three times greater than those found using the standard single trait GWAS, and it nearly doubled the number of per trait associations for another trait.

2.2 Results

2.2.1 Methods overview

2.2.1.1 Pleiotropic association test

Our method, PAT (pleiotropic association test), takes in GWAS summary statistics measured for T traits and assumes each variant is drawn according to the multivariate normal (MVN) distribution: $S \sim \mathcal{N}(0, \Sigma)$. Furthermore, it assumes the covariance matrix can be decomposed into two independent components, environment and genetics ($\Sigma = \Sigma_e + \Sigma_g$). With this assumption in mind, PAT performs a likelihood ratio test (LRT) between two proposed MVN distributions. The null hypothesis is $\Sigma_g = 0$; therefore, the summary statistics for one variant, $S = \{s_1, ..., s_T\}$ has the following distribution: $S \sim \mathcal{N}(0, \Sigma_e)$.

Under the alternative hypothesis ($\Sigma_g \neq 0$), PAT models the genetic effect size according to the polygenic model and assumes the standard genetic correlation structure between traits [6, 118, 153]. This results in summary statistics having the following distribution: $S \sim \mathcal{N}(0, \Sigma_g + \Sigma_e)$.

Having now defined the distributions, a LRT can be computed for each variant's set of summary statistics S. Using the critical value κ for the threshold of significance, it can now be decided whether a variant is associated with the set of traits.

$$\frac{P(S|\mu = 0, \Sigma = \Sigma_e + \Sigma_g)}{P(S|\mu = 0, \Sigma = \Sigma_e)} > \kappa \tag{2.1}$$

While likelihood ratio tests approximately follow a mixture of χ^2 distributions, utilizing a χ^2 distributions can be complicated and may have reduced power [155]. Therefore, instead of a closed form solution, PAT efficiently uses null simulations to determine significance (see Methods). Additionally, we note that when there is no environmental correlation ($\Sigma_e = I$), PAT is comparable to a Wald test.

2.2.1.2 Multi-Trait GWAS interpretation

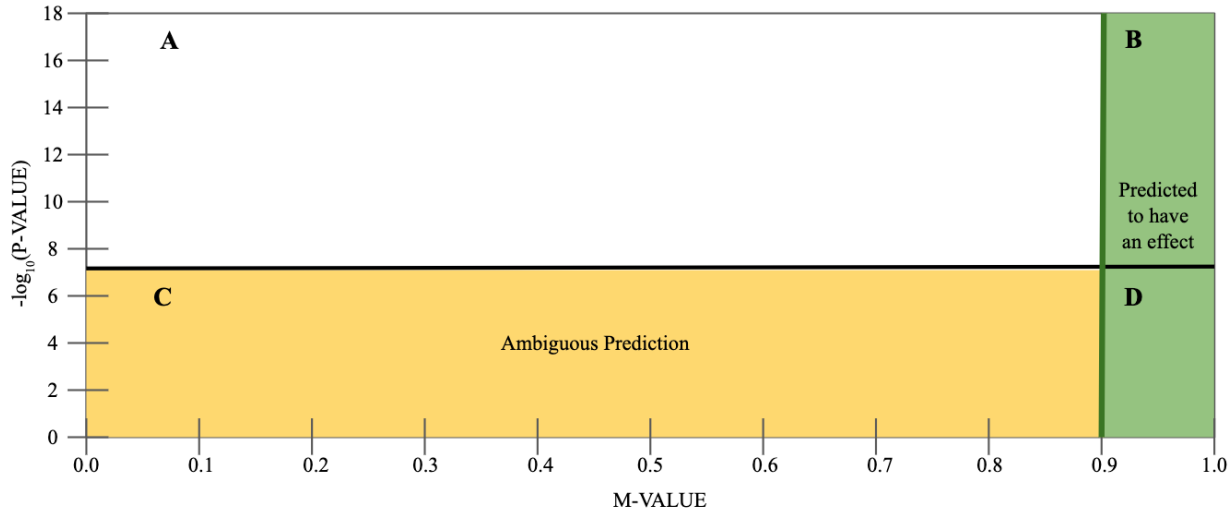


Figure 2.1: **Interpreting m-values using a P-M plot.** Along the x-axis is the per trait m-value and the y-axis shows the p-value from the original single trait GWAS. Region A is when the original association is significant, but the m-value interpretation is ambiguous. There should not be data points in this region. Region B and D are associations with an m-value greater than 0.9, so the interpretation is that there is a genetic effect in this trait. In Region C, the m-value interpretation is left ambiguous.

While PAT is a powerful tool for testing multi-trait associations, it is an omnibus test and only provides one p-value per variant. As a result, even when the null is rejected, we lack clarity as to which trait(s) drive the association; therefore, we propose a per trait p-value interpretation by estimating the posterior probability of a variant having a non-zero effect on a trait. This framework, m-values, was originally developed for interpreting meta-analysis across studies, but here it is extended to account for the covariance structure between traits [63].

To provide some intuition on m-values, we will describe the P-M plot (p-value by m-value plot) [63] in Fig 2.1. This plot has the p-value from the original single trait GWAS along the y-axis and the corresponding m-value along the x-axis. A line at $-log(5 \times 10^{-8})$ denotes the threshold where a variant is considered genome-wide significant. Region A is where the original single trait GWAS resulted in the variant being significant while the interpretation of the omnibus test did not. There should not be data points in this region.

Regions B and D contain variants interpreted as associated with the trait because the m-value is greater than 0.9. Some of these variants have already been identified by the single trait GWAS (B) while other traits will be uniquely discovered on a per trait level (D). Region C contains the variants whose m-value is less than or equal to 0.9 and were left with an ambiguous interpretation.

2.2.2 Covariance structure between traits impacts PAT's rejection region

We now present an overview of PAT and its rejection region by comparing its shape to the rejection region of a version of standard GWAS generalized to multiple traits called multiple independent GWAS (MI GWAS). We chose to compare to this method over standard single trait GWAS because it accounts for multiple testing while being less stringent than a Bonferroni correction. MI GWAS works by testing if the largest summary statistic per trait was larger than the critical value for significance set using null simulations.

In Fig 2.2, we simulated 100,000 summary statistics for two traits with the genetic variance ($\sigma_g^2 = 4.9 \times 10^{-5}$) and the sample size (N=25,000) equal for both traits; the level of significance was $\alpha = 0.05$. The first column highlights MI GWAS's performance. Variants which were correctly identified as associated are shown in red while the ones missed by MI GWAS are grey. In each row regardless of model specification, the shape of MI GWAS's rejection region was a square. As MI GWAS does not account for genetic correlation, there was no effect on the critical value when this parameter varied.

The same phenomenon was not true for PAT which is shown in the second column (with the critical values of MI GWAS depicted with black lines). Here, when PAT rejected the null hypothesis, the data points are in blue, and those PAT failed to reject are grey. In all four rows the shape of the rejection region was an ellipse. As PAT models environmental and genetic correlation, both parameters impacted the shape of the elliptical rejection region. In the first row there is no environmental or genetic correlation, so the shape was exactly a circle. This means any extreme value for at least one of the summary statistics was likely to be rejected. In the second row, we modeled 67% genetic correlation and no environmental

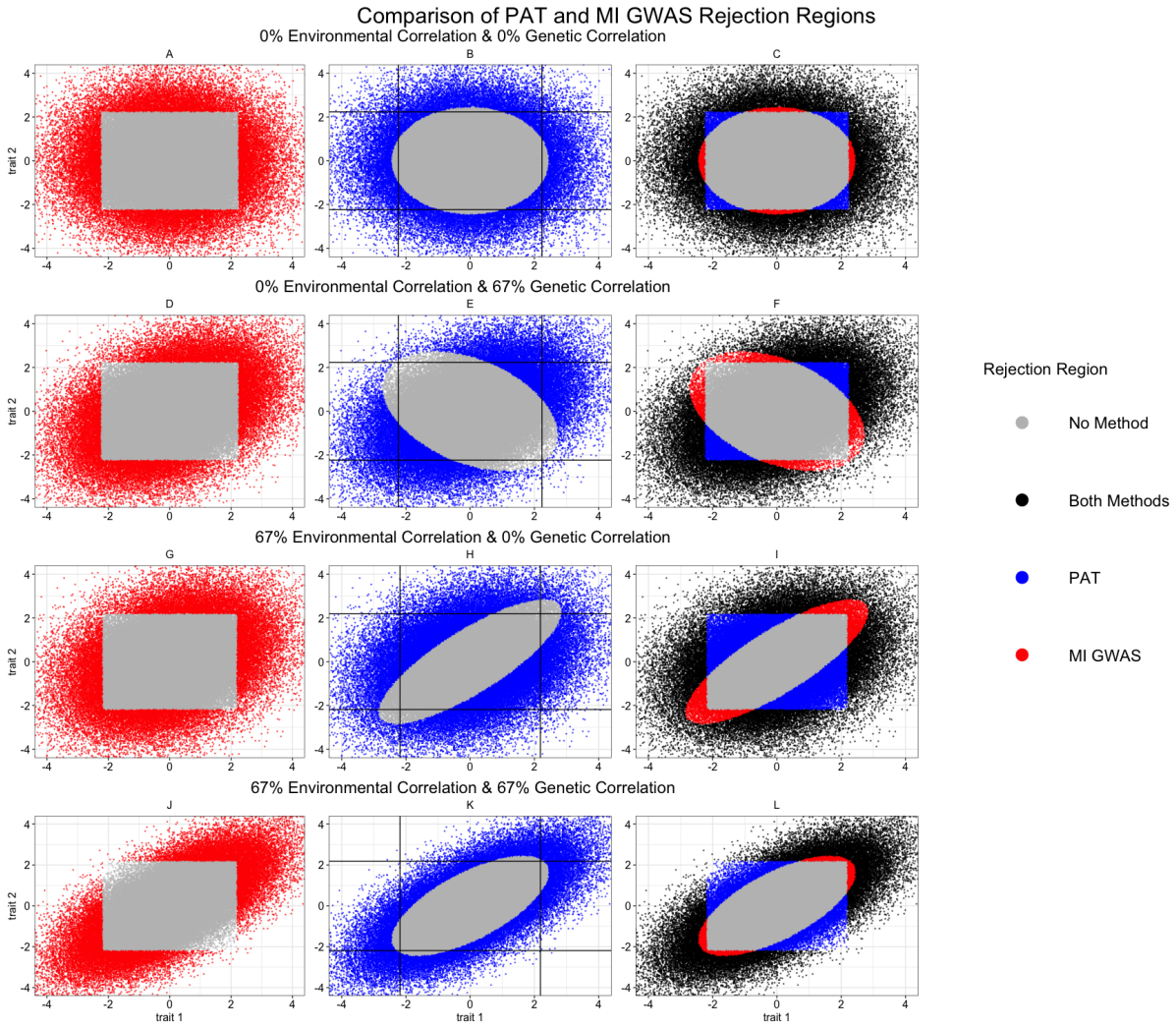


Figure 2.2: **Comparison of the rejection regions for MI GWAS and PAT.** We simulated 100,000 summary statistics for two traits with the genetic variance ($\sigma_g^2 = 4.9 \times 10^{-5}$) and the sample size (N=25,000) equal for both traits. We varied the genetic and environmental correlation between traits and used $\alpha = 0.05$ for the level of significance. Each row corresponds to one set of simulations highlighting three points. The left column shows the rejection region of MI GWAS, the middle column has PAT's rejection region while the third column provides a comparison of the two methods. The simulations used in sub-figures A-C have no environmental or genetic correlation while the data in sub-figures D-F has no environmental correlation and 67% genetic correlation. For the third row of sub-figures, G-I, the environmental correlation was 67% while there was no genetic correlation between traits. The last row of simulations assumed an environmental and genetic correlation of 67%.

correlation. Here, the shape enabled PAT to correctly identify more variants with positively correlated z-scores but failed to aid in identifying variants with negatively correlated z-scores. This follows the intuition that modeling genetic correlation would increase power to identify

variants whose summary statistics followed this correlation pattern.

While the first two rows followed intuition, the shape of PAT's rejection region in the last two rows was less intuitive. In the third row, we simulated traits with 67% environmental correlation but no genetic correlation. In this situation, the shape of the rejection region was in the direction of the environmental correlation; therefore, PAT has more power when z-scores were negatively correlated relative to when they were positively correlated. This means when summary statistics were positively correlated, PAT failed to reject the null unless the values were very extreme because it assumed the only source of positive correlation was the environment. The gain in power in the direction of negative correlation was due to the same idea that these values were unlikely under a positively correlated environment unless there was a non-environmental effect (i.e., genetics). In the final row, we simulated a positively correlated environment and genetics. Here, the shape of the rejection region still followed the direction of the environmental correlation. This aided in controlling false positives, but it meant that PAT may have been overly conservative in the direction of environmental correlation even when there was genetic correlation in the same direction. Further to that point, the critical value for MI GWAS as shown in sub-figures H and K (black lines) was identical. In sub-figure K, there were fewer variants pass this cut-off that were missed by PAT relative to sub-figure H. This means that while PAT was consistently more conservative in the direction of the environmental correlation, it was less conservative when it expected a genetic reason for correlated summary statistics.

The right most column has a comparison of the relative power of PAT and MI GWAS. Variants that were correctly identified by both methods are black data points while those missed by both are grey. The variants only identified as significant by PAT are blue while those found only by MI GWAS are red. Under all four simulation frameworks, PAT had more statistical power than MI GWAS with the greatest improvement occurring when there was genetic correlation and no environmental correlation (sub-figure F) . We note that the blue region may appear smaller in sub-figure F than in sub-figures I and L, but the density of the data points was higher.

2.2.3 M-values provide accurate interpretation of omnibus association tests

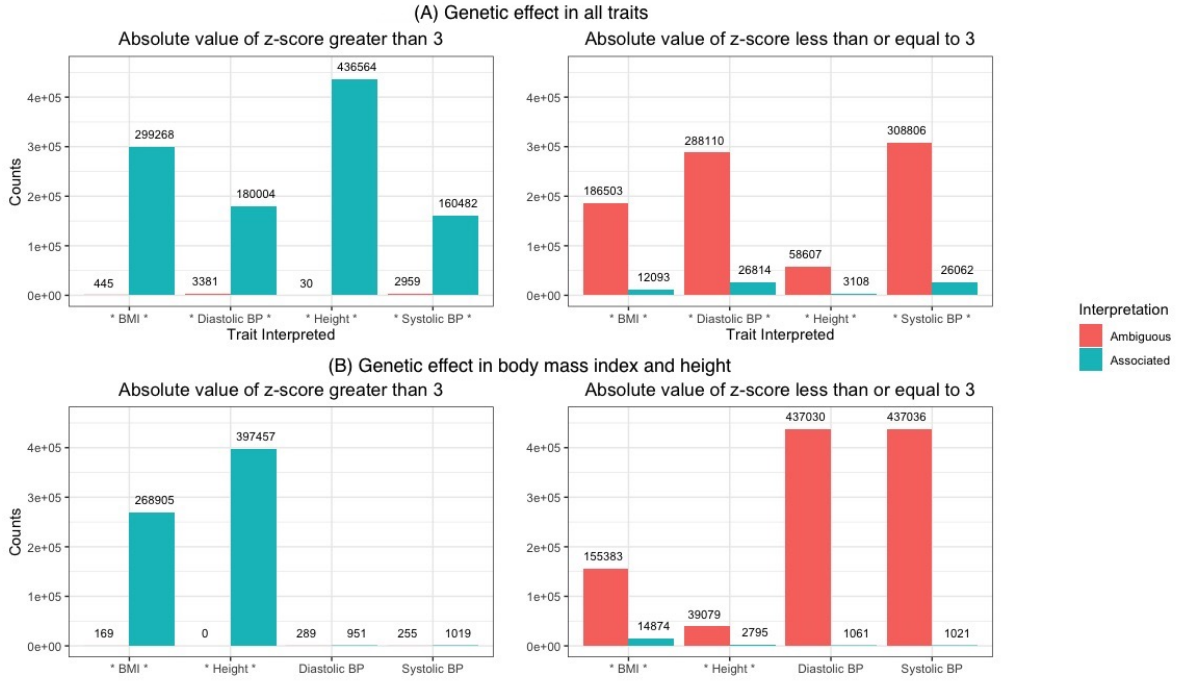


Figure 2.3: **Interpreting per trait associations from omnibus significant variants.** We simulated one million variants for four traits under two models. The first set of simulations assumed there was a genetic effect in every trait (A), while the second model only has a genetic effect in body mass index and height (B). The associated traits are noted with an asterisks (*). The results for each trait were split based on the absolute value of the z-score and showed the interpretation as either ambiguous or associated. The threshold for associated is an m-value greater than 0.9.

In section 2.2.2, we provided an overview of how PAT provides a per variant omnibus p-value for the set of traits. Here, we used simulated data reflective of four real UK Biobank traits (see Methods) to provide some intuition about m-values as well as highlight its accuracy. M-values were produced by enumerating over the set of configurations $C = \{0, 1\}^4$ which indicate which trait(s) have a genetic effect, $\Sigma_g(c)$. We note that the configuration $c = (1, 1, 1, 1)$ indicates a genetic effect in all four traits (i.e. $\Sigma_g(c) = \Sigma_g$). For each configuration, we calculated the posterior probability $P(S|\mu = 0, \Sigma = \Sigma_e + \Sigma_{g(c)})$. We then take the sum of the configurations compatible with trait i ($c_i = 1$) and divide by the total probability over all configurations to produce the m-value for trait i. If this ratio $m_i > 0.9$,

we interpreted the omnibus variant-trait association to be an association between the variant and trait i. If this ratio $m_i \leq 0.9$, we left the interpretation as ambiguous. We note that m-values are a Bayesian quantity whose threshold is a matter of convention established in previous work [63].

In Fig 2.3, we simulated one million variants under two model conditions. In sub-figure A, there was a genetic effect in all four traits: body mass index, diastolic blood pressure, height, and systolic blood pressure while in sub-figure B, we modeled a genetic effect in only body mass index and height. In Fig 2.3, the truly associated traits were denoted with an asterisks (*) around the trait name. The effect sizes were simulated such that the first model has 50% power and 44% when there was a genetic effect in only body mass index and height. We split the summary statistics (z-scores) for each trait based on whether there was even modest signal in a particular trait ($|\text{z-score}| > 3$). This distinction was due to differing expectations on the ability to correctly interpret an association. We note that the inclusion of variants with a $|\text{z-score}| \leq 3$ for a particular trait was primarily done for completeness and their interpretation was overwhelmingly ambiguous (Fig 2.3 right panel). We therefore, focus on the left panel of Fig 2.3 where the $|\text{z-score}| > 3$.

When there was a genetic effect in all four traits (top row left side), the m-value was greater than 0.9 for the vast majority of z-scores which means the majority of variants were correctly interpreted as associated with all traits. Diastolic and systolic blood pressure had the most ambiguous associated variants with 3,381 and 2,959, respectively. This, however, was still less than 2% of the variants with at least a modest effect size ($|\text{z-score}| > 3$) being interpreted as ambiguous for each trait. Furthermore, when there was a modest effect size the overall false negative rate was 0.6% across the traits.

The second set of simulations modeled a genetic effect in only two of the traits (bottom row left side). Here, the m-value framework correctly interpreted when there was a genetic effect in body mass index and height for most significant variants. For body mass index only 169 of the variants were missed and none were left ambiguous for height. For diastolic blood pressure and systolic blood pressure, approximately 1,200 variants for each trait that had a

|z-score| > 3. The m-value wrongly identified 951 and 1,019 of those variants as associated for diastolic blood pressure and systolic blood pressure, respectively. Overall, 99.5% of the variants analyzed for diastolic blood pressure and systolic blood pressure were left with an ambiguous interpretation. For body mass index and height, 64.5% and 90.5% of all variants, respectively, were correctly interpreted as associated when there was only a genetic effect in these two traits.

These simulations show the m-value framework has a low false positive assignment rate, and enabled the correct classification of many associated variants. This was especially true when |z-score| > 3 while |z-score| $\leq$ 3 typically resulted in the interpretation being ambiguous regardless of the ground truth. While this was still a false negative assignment, many of these associations would have failed to pass a nominal test for significance (p-value < 0.05).

2.2.4 PAT is a powerful omnibus method for multi-trait GWAS

Now that we have established the intuition behind PAT, it is important to understand its performance relative to other multi-trait methods. Here, we compare four methods: PAT, MTAG, HIPO, and ASSET [14, 126, 151]. HIPO is an omnibus method that performs eigenvalue decomposition resulting in orthogonal components each of which is used to create a weighted sum of z-scores. For this comparison, z-scores from all components are considered simultaneously and a variant is deemed associated as long as it is genome-wide significant for at least one component. Another method is MTAG, and it also uses a weighted sum of z-scores. MTAG, however, is not an omnibus method but tests each trait separately while leveraging information from the other traits. The results from MTAG are converted to an omnibus test by determining if the variant is genome-wide significant for at least one trait. The final method is ASSET; this method works by searching for the subset of traits with the strongest positive signal and separately the strongest negative signal. ASSET then combines these test statistics using a chi-squared method to form an overall test statistic which we use for comparison. While MTAG and HIPO generate multiple test statistics for each variant, we do not correct for multiple testing; all methods are tested at $\alpha = 5 \times 10^{-8}$.

Genetic Effect	Number of Variants	$\frac{\Sigma_{h2}}{\text{causal}}$	Genome-Wide Significant			
			PAT	HIPO	MTAG	ASSET
No Trait	1,350,000	0	0	0	0	0
$\mathcal{B}, \mathcal{D}, \mathcal{H}, \mathcal{S}$	5,000	40,000	**113**	54	60	103
	3,000	24,000	**198**	113	119	196
	2,000	16,000	291	194	196	**326**
$\mathcal{B}, \mathcal{D}, \mathcal{H}$	5,000	40,000	**108**	53	56	76
	3,000	24,000	**204**	113	121	170
	2,000	16,000	**307**	216	226	286
$\mathcal{B}, \mathcal{D}, \mathcal{S}$	5,000	40,000	0	2	**3**	1
	3,000	24,000	0	**12**	7	5
	2,000	16,000	0	**26**	12	11
$\mathcal{B}, \mathcal{H}, \mathcal{S}$	5,000	40,000	**124**	73	58	92
	3,000	24,000	**216**	166	128	199
	2,000	16,000	**352**	281	219	334
$\mathcal{D}, \mathcal{H}, \mathcal{S}$	5,000	40,000	**88**	28	36	56
	3,000	24,000	**161**	105	111	131
	2,000	16,000	**257**	173	176	227
$\mathcal{B}, \mathcal{D}$	5,000	40,000	0	**2**	1	0
	3,000	24,000	0	**15**	6	2
	2,000	16,000	0	**34**	4	1
$\mathcal{B}, \mathcal{H}$	5,000	40,000	**96**	36	40	60
	3,000	24,000	**160**	106	116	138
	2,000	16,000	**260**	196	201	255
$\mathcal{B}, \mathcal{S}$	5,000	40,000	0	**33**	11	5
	3,000	24,000	5	**81**	32	30
	2,000	16,000	12	**128**	61	48
$\mathcal{D}, \mathcal{H}$	5,000	40,000	**90**	40	42	41
	3,000	24,000	**177**	111	114	127
	2,000	16,000	**253**	195	185	204
$\mathcal{D}, \mathcal{S}$	5,000	40,000	0	**3**	0	0
	3,000	24,000	0	**7**	2	2
	2,000	16,000	0	**23**	14	9
$\mathcal{H}, \mathcal{S}$	5,000	40,000	**80**	40	32	46
	3,000	24,000	**185**	127	94	131
	2,000	16,000	**225**	191	144	179
$\mathcal{B}$	5,000	40,000	0	**12**	8	4
	3,000	24,000	1	**32**	20	9
	2,000	16,000	6	**51**	45	30
$\mathcal{D}$	5,000	40,000	0	**5**	1	0
	3,000	24,000	0	**14**	4	1
	2,000	16,000	1	**35**	15	7
$\mathcal{H}$	5,000	40,000	**89**	36	46	47
	3,000	24,000	**154**	82	92	94
	2,000	16,000	**191**	126	139	134
$\mathcal{S}$	5,000	40,000	0	**11**	1	0
	3,000	24,000	0	**39**	3	2
	2,000	16,000	1	**66**	4	2
Total	1,500,000	—	**4,405**	3,486	3,005	3,820
Running Time (seconds)	—	—	72	96	150	54,709

Table 2.1: **Comparison of multi-trait GWAS methods.** 1.5 million variants were simulated with z-scores for four traits with 10% of variants as truly associated. The first column lists which trait has a genetic effect. The second column is the number of variants simulated under this specific model. The third column is the genetic effect size. The remaining four columns contain the number of variants identified as associated by four methods: PAT, HIPO, MTAG, and ASSET. The final row of the table contains each methods running time.

In Table 2.1, 1.5 million z-scores were simulated for four traits with the environmental and genetic covariance structure based on four traits from the UK Biobank (see Methods) and 10% (150,000) of the variants were causal in at least one trait [111, 147]. The first row in Table 2.1 corresponds to the 1,350,000 variants simulated under the null. All four methods correctly identified zero associated variants. The remaining 150,000 truly associated variants were equally split across all configurations of genetic effect. For each of the configurations, there were three scaling factors for the heritability covariance matrix, Σ_{h2}, which can be thought of as the number of causal variants ($\frac{\Sigma_{h2}}{\text{causal}}$) where causal equals 40k, 24k, or 16k. We note that the methods assume the polygenic model (i.e. $\Sigma_g = \frac{\Sigma_{h2}}{\#\text{ variants}}$), but we simulated assuming fewer causal variants to create effect sizes large enough that there was power for discovery. The 10,000 variants for each configuration were split such that 5k, 3k, and 2k simulations came from each of the respective causal effect sizes. The configurations are subsets of the four traits, body mass index ($\mathcal{B}$), diastolic blood pressure ($\mathcal{D}$), height ($\mathcal{H}$), and systolic blood pressure ($\mathcal{S}$). The final row in Table 2.1 contains the running time for each method. Here, we see ASSET was significantly slower than the other three methods which were comparable to each other. PAT's efficient running time indicates that the use of importance sampling can enable a speed up comparable to deriving p-values analytically; the differences in compute time between PAT, HIPO, and MTAG were likely due to other factors (e.g. MTAG does a number of sanity checks prior to analysis).

While no simulation framework truly reflects the real world, this arrangement attempted to non-exhaustively model different scenarios that occur when analyzing z-scores from multiple traits. Namely, we explored the power to discover summary statistics with different causal effect sizes and violations of a pleiotropic effect in all traits. Under the various configurations shown here, all of the methods were under powered due to the simulations being centered around zero; however, PAT was the most powerful method in nearly half of the simulated scenarios as well as overall. Across all scenarios PAT identified 4,405 associated variants which was an 15.3% increase over ASSET (3,820), a 26.4% increase over HIPO (3,486) and a 46.6% increase over MTAG (3,005). While PAT generally performed the best, the other methods

did significantly better when the genetic effect in height was absent. Without considering environmental correlation, this scenario was similar to that seen in Fig 2.2 sub-figures (B) and (E). There we saw that the closer one trait's z-score was to 0, the larger the other trait's effect size needed to be. Another factor was the environmental correlation; the other three traits have more environmental correlation to each other than to height which was similar to the scenario in Fig 2.2 sub-figure (K). In this case, PAT was shown to be conservative in the direction of environmental correlation. Finally, we explore the simulations from this section on a per trait level in section 2.2.5.

2.2.5 M-values enable more per trait interpretations in multi-trait GWAS

The four multi-trait methods were previously compared in regards to their power to perform omnibus association testing (see section 2.2.4). Here, we investigated the per trait interpretation of these associations. As MTAG computes a p-value for every trait, the method provides a direct per trait interpretation; therefore, for each respective trait we reported the variants with a p-value $< 5 \times 10^{-8}$. The method, ASSET, considers all possible subsets and selects the one that maximizes its test statistic. This is done separately in the positive and negative directions of effect and are then combined for a two-tailed test which determines the omnibus association. For the associated variants, we tested each direction separately for significance (p-value $< 5 \times 10^{-8}$ and interpreted the subsets that produced a significant association as the trait(s) driving the association. The last two methods, HIPO and PAT, only provided an omnibus interpretation; therefore, we applied the m-value framework to assign a per trait association to variants whose omnibus p-value $< 5 \times 10^{-8}$. For both methods, this was done by taking the associated variants and calculating the posterior predictive probability (m-value) of whether there was a genetic effect in each particular trait. If the m-value was greater than 0.9, the variant was deemed associated with the trait. Otherwise, the interpretation was left ambiguous.

Prior to exploring the per trait interpretation, we note that only MTAG controls the false positive per trait interpretation due to its use of p-values for the assignment; m-values do not

Genetic Effect	Number of Variants	$\frac{\Sigma_{h2}}{\text{causal}}$	Body Mass Index				Diastolic Blood Pressure				Height				Systolic Blood Pressure			
			PAT	HIPO	MTAG	ASSET	PAT	HIPO	MTAG	ASSET	PAT	HIPO	MTAG	ASSET	PAT	HIPO	MTAG	ASSET
	5,000	40,000	**28**	17	10	15	**5**	3	0	4	**113**	44	50	74	**8**	2	0	1
$\mathcal{B},\mathcal{D},\mathcal{H},\mathcal{S}$	3,000	24,000	**64**	38	17	39	**31**	15	2	10	**198**	96	98	143	**35**	22	3	16
	2,000	16,000	**131**	89	44	78	**71**	44	8	21	**284**	157	139	221	**70**	56	18	47
	5,000	40,000	**34**	15	8	12	**11**	10	1	4	**108**	43	47	56	1	1	0	1
$\mathcal{B},\mathcal{D},\mathcal{H}$	3,000	24,000	**64**	42	26	43	**41**	22	7	16	**200**	87	89	118	2	1	0	1
	2,000	16,000	**139**	96	47	85	**85**	61	18	29	**301**	182	171	224	2	3	0	0
	5,000	40,000	0	0	0	0	0	**1**	**1**	0	0	0	0	0	0	**2**	**2**	1
$\mathcal{B},\mathcal{D},\mathcal{S}$	3,000	24,000	0	0	0	0	0	**8**	5	3	0	0	0	0	0	**10**	3	5
	2,000	16,000	0	0	0	0	0	**15**	7	9	0	0	0	0	0	**21**	5	6
	5,000	40,000	**39**	29	12	12	3	2	0	1	**124**	52	46	64	16	**25**	0	1
$\mathcal{B},\mathcal{H},\mathcal{S}$	3,000	24,000	**74**	60	25	49	3	1	0	0	**215**	108	101	145	51	**63**	3	17
	2,000	16,000	**160**	125	53	92	4	3	0	0	**345**	206	170	244	97	**115**	6	40
	5,000	40,000	1	1	0	0	**6**	1	1	4	**88**	26	35	46	**8**	3	0	6
$\mathcal{D},\mathcal{H},\mathcal{S}$	3,000	24,000	1	1	0	0	**24**	19	3	6	**161**	98	107	116	**25**	18	3	6
	2,000	16,000	3	1	0	1	**53**	37	9	11	**257**	162	163	195	**61**	40	8	21
	5,000	40,000	0	0	0	0	0	**1**	**1**	0	0	0	0	0	0	1	0	0
$\mathcal{B},\mathcal{D}$	3,000	24,000	0	0	0	0	0	**14**	6	2	0	0	0	0	0	0	0	0
	2,000	16,000	0	0	0	0	0	**32**	4	1	0	1	0	0	0	0	0	0
	5,000	40,000	**26**	14	6	15	2	0	0	1	**96**	31	34	50	1	0	0	1
$\mathcal{B},\mathcal{H}$	3,000	24,000	**57**	44	28	29	0	0	0	1	**159**	88	90	110	2	1	0	1
	2,000	16,000	**116**	102	64	76	0	0	0	0	**255**	151	144	177	1	1	0	0
	5,000	40,000	0	**19**	11	5	0	1	0	0	0	0	0	0	0	**18**	0	1
$\mathcal{B},\mathcal{S}$	3,000	24,000	5	**47**	30	26	0	0	0	0	0	1	0	1	3	**56**	2	3
	2,000	16,000	10	**85**	54	41	0	0	0	0	0	0	0	2	6	**74**	7	11
	5,000	40,000	1	0	0	0	**7**	5	0	0	**90**	37	42	40	0	1	0	0
$\mathcal{D},\mathcal{H}$	3,000	24,000	3	0	0	2	**36**	23	3	4	**177**	100	111	117	0	0	0	0
	2,000	16,000	3	2	0	3	**87**	81	18	24	**253**	174	173	185	3	2	0	0
	5,000	40,000	0	0	0	0	0	**1**	0	0	0	0	0	0	0	**2**	0	0
$\mathcal{D},\mathcal{S}$	3,000	24,000	0	0	0	0	0	**4**	1	2	0	0	0	1	0	**4**	2	2
	2,000	16,000	0	0	0	0	0	**14**	8	7	0	0	0	0	0	**17**	6	5
	5,000	40,000	1	0	0	0	1	2	0	0	**80**	33	32	40	**12**	10	0	3
$\mathcal{H},\mathcal{S}$	3,000	24,000	1	0	0	1	2	1	0	1	**185**	99	92	119	39	**48**	2	11
	2,000	16,000	5	2	0	6	2	1	0	0	**225**	155	141	166	54	**83**	3	17
	5,000	40,000	0	**12**	8	4	0	0	0	0	0	0	0	0	0	0	0	0
$\mathcal{B}$	3,000	24,000	1	**32**	20	9	0	0	0	0	0	0	0	0	0	0	0	0
	2,000	16,000	6	**51**	45	30	0	0	0	0	0	0	0	1	0	1	0	0
	5,000	40,000	0	1	0	0	0	**4**	1	0	0	0	0	0	0	0	0	0
$\mathcal{D}$	3,000	24,000	0	0	0	0	0	**14**	4	1	0	0	0	0	0	0	0	0
	2,000	16,000	0	0	0	0	1	**33**	15	6	0	0	0	0	0	1	0	0
	5,000	40,000	1	1	0	0	1	1	0	0	**89**	36	46	47	1	1	0	0
$\mathcal{H}$	3,000	24,000	2	1	0	1	1	1	0	0	**154**	82	92	94	1	1	0	1
	2,000	16,000	0	0	0	0	2	1	0	0	**191**	126	139	133	1	1	0	0
	5,000	40,000	0	0	0	0	0	0	0	0	0	0	0	0	0	**11**	1	0
$\mathcal{S}$	3,000	24,000	0	0	0	0	0	0	0	0	0	0	0	0	0	**39**	3	2
	2,000	16,000	0	0	0	0	0	0	0	0	0	0	0	0	1	**66**	4	2

Table 2.2: **Four multi-trait GWAS methods with per trait interpretation.** 1.5 million variants were simulated with z-scores for four traits with 10% of variants being truly associated. The first column lists which trait has a genetic effect. The second column is the number of variants simulated under this specific model. The third column is the genetic effect size of the variant. The remaining columns are split by trait where the performance of the four methods are shown for each trait. These 16 columns present the number of variants identified as associated by each method for the specific trait. MTAG uses p-values, ASSET uses the optimal subset, while PAT and HIPO use the m-value framework to provide per trait associations.

directly control for false positives. As a result, m-values are only meant to provide empirical insights and interpretation to p-values not replace them. This means the comparisons in Table 2.2 between MTAG's p-values and the m-value interpretations are not an apples to apples comparison. In Fig 2.4, we provide a fairer comparison by ranking the p-values and m-values (see section 2.2.6). There we show that for any false positive rate, PAT and HIPO have more true positive per trait assignments than MTAG. Separately, we acknowledge

that while ASSET provides the subset of traits with the strongest association signal with the intent of a more interpretable multi-trait association. It is possible that a trait was included in the optimal subset due to its tagging the causal signal in another trait. In this case, including the trait was useful for increasing the association power but would lead to a spurious interpretation.

In Table 2.2, all methods analyzed 1.5 millions simulations with 10% (150,000) causal variants equally divided across all configurations of genetic effect. For each of the configurations, different effect sizes were also considered (see section 2.2.4). In Table 2.2, the number of per trait associations was reported by trait under each configuration. When the variant did not truly have a genetic effect on the trait, the box was greyed to indicate false positives. Overall, Table 2.2 resembled the results shown in Table 2.1. One example of an exception was when there was a genetic effect in body mass index, height, and systolic blood pressure $(\mathcal{B}, \mathcal{H}, \mathcal{S})$. While PAT identified more associated variants, HIPO has more per trait associations for systolic blood pressure. This means that while HIPO has less power than PAT for the omnibus test (see Table 2.1), it was able to provide the most per trait interpretations for this trait. This was due to HIPO identifying different associated variants than PAT which were then interpreted on a per trait level. We also saw this phenomenon when there was a genetic effect in height and systolic blood pressure $(\mathcal{H}, \mathcal{S})$.

Overall, PAT identified 6,264 true per trait associations from its 4,405 omnibus associations. For HIPO, the m-value framework interprets 4,557 true per trait associations from its 3,486 significant variants. When comparing PAT to HIPO, there are 37.5% more true per trait associations than HIPO due to PAT having more power as an omnibus method. The method, ASSET, identified 3,820 significant associations with 3,944 traits correctly placed in the optimal subset. Finally, we consider MTAG which directly identified 3,064 total per trait associations (3,005 omnibus associations). While HIPO and PAT identified 16.0% and 46.6% more omnibus associations than MTAG, respectively the m-value framework enabled a 48.7% increase for HIPO and a 104.4% increase for PAT in per trait associations relative to MTAG, a method designed for per trait interpretation. When comparing HIPO and

PAT and their m-values to ASSET, we saw that HIPO had 8.7% fewer omnibus associations than ASSET but 15.5% more per trait assignments. Separately, while PAT had 15.3% more omnibus associations than ASSET, there were 58.8% more per trait findings.

While m-values enabled a significant increase in per trait interpretations, as stated before, the m-value threshold does not directly control for false positives. In Table 2.2, MTAG had no false positive per trait associations. The m-values produced for PAT and HIPO, however, did result in a small number of false positive assignments, 58 and 42 respectively. This was 0.92% and 0.91% of their respective per trait interpretations. When we considered the subsets produced by ASSET, we observed there were 28 false positive placements (0.73%).

2.2.6 M-values produce a higher true positive rate than MTAG

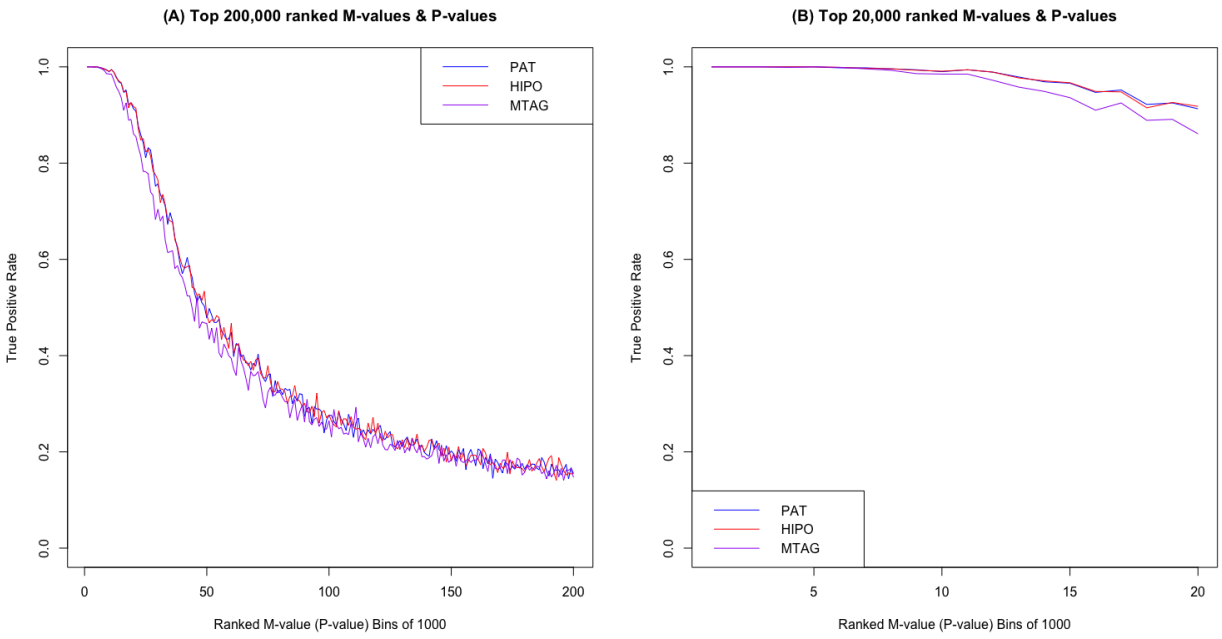


Figure 2.4: **Comparison of m-values and p-values.** M-values were assigned to all z-scores for PAT and HIPO. For each method, they were ranked and placed in bins of 1,000. The p-values from MTAG were also ranked and binned in sets of 1,000. A comparison of their respective true positives rates are shown in (A) the first 200 bins and (B) first 20 bins.

We previously compared the per trait m-values produced for PAT and HIPO to the per trait p-values produced by MTAG. In Table 2.2, we observed that through m-values

PAT identified 6,264 true per trait associations and HIPO discovered 4,557 true per trait associations. MTAG directly provided a per trait interpretation for a total of 3,064. While these results indicated PAT was the most powerful approach, they also provided evidence that computing posterior predictions (m-values) after omnibus associations was a more powerful approach to association testing than directly analyzing each trait using MTAG. While the number of true positives in section 2.2.5 supported this claim, the difference in the false positive rate between m-values and p-values draw this claim into question. This is to say, if MTAG was allowed to produce the same number of false positives as m-values, it is possible that MTAG would be the most powerful approach.

In order to test this claim, m-values must be modified to better reflect p-values. Currently, m-values were only generated for variants deemed genome-wide significant. This was due to their design as an interpretation framework. M-values were not designed to replace p-values but to elucidate what traits may be driving the significant association. In this comparison, we assigned m-values for every p-value instead of only to the omnibus significant ones. This enabled a comparison of all posterior predictions to the per trait p-values of MTAG. While illustrative of the power of m-values, this approach is not advisable in practice. M-values should only be used as a means of interpreting omnibus associations.

For this comparison, we used the 1.5 million simulations with 10% causal variants previously described in section 2.2.4. As stated previously, MTAG directly produced a per trait p-value and m-values were assigned to HIPO and PAT. In Fig 2.4, we ranked the m-values for PAT and HIPO and the p-values for MTAG. For p-values, the rank order was from $[0.0, 1.0]$ while m-values were ordered from $[1.0, 0.0]$. For each m-value (and p-value), whether or not the variant was truly associated with the trait was known; therefore, the true positive rate for the variants in rank order could be calculated. After ranking the m-values (and p-values), the variants were binned in sets of 1,000 for each method. In Fig 2.4, sub-figure (A), the first 200 bins or 200,000 top p-values (m-values) for each method are along the x-axis. Along the y-axis, the true positive rate for each bin is shown. For the most significant associations by p-values and m-values the true positive rate was 1.00. As the ranked position decreased,

the true positive rate decreased. Overall, the general trend of m-values for both HIPO and PAT and p-values for MTAG followed the same pattern. In the top left of sub-figure (A) MTAG (purple) has a slightly lower true positive rate than PAT (blue) and HIPO (red) which overlay each other.

In Fig 2.4 sub-figure (B), we explored the top 20 bins or 20,000 m-values and p-values more closely. Here, we saw that there was some separation between the m-values for PAT and HIPO, respectively shown in blue and red, and MTAG in purple. From this, there is evidence that while the m-value framework did not control for false positives directly, it does have increased power relative to a directly interpretable multi-trait method, such as MTAG. We note that while true, neither the true (nor false) positive rate can be elucidated from the m-value directly. Therefore, m-values should only be used as designed to interpret significant p-values.

2.2.7 PAT discovers novel per trait associations in the UK Biobank

Trait Interpreted	Directly From GWAS		M-value >0.90		
	Single Trait GWAS	MTAG	MI GWAS	HIPO	PAT
body mass index ($\mathcal{B}$)	37,205	32,527	64,706	65,462	**67,139**
diastolic blood pressure ($\mathcal{D}$)	18,593	17,610	56,369	**58,294**	56,271
height ($\mathcal{H}$)	160,227	117,882	155,730	136,519	**191,420**
systolic blood pressure ($\mathcal{S}$)	17,515	16,927	48,308	**51,234**	49,125
Total	233,540	184,946	325,113	311,509	**363,955**

Table 2.3: **UK Biobank data interpretation.** We analyzed four traits from the UK Biobank using five methods: Single Trait GWAS, MTAG, MI GWAS, HIPO, and PAT and show the variants associated with each trait. For Single Trait GWAS and MTAG, the per trait association was directly computed. For MI GWAS, HIPO and PAT, an omnibus association was first performed. The significant variants were then interpreted using the m-value framework using 0.9 as the threshold.

While simulations have indicated PAT is a powerful method for association testing and m-values enable a per trait interpretation, we now apply this two step approach to real data. We analyzed the UK Biobank summary statistic for body mass index, diastolic blood

pressure, height, and systolic blood pressure (see Methods) [111, 147]. Here, five methods were compared: Single Trait GWAS (how the z-scores and p-values were derived), MTAG, MI GWAS, HIPO and PAT [126, 151]. The set of variants were processed such that only variants which were biallelic, have non-ambiguous strands, a minor allele frequency greater than 1%, and an INFO score greater than 80% were retained. This left 7,025,734 variants that meet the criteria for all four traits. The reference and alternate allele were coordinated across traits by flipping the direction of the effect when necessary. LD-Score regression and cross-trait LD-Score regression were used to calculate the genetic and environmental covariance structure [17, 18]

Using standard single trait GWAS, there were 211,546 uniquely associated variants across the four traits of interest. With MTAG, 164,263 uniquely associated variants were identified, 931 of which were novel associations. MI GWAS implicated 183,669 variants as associated, but none of the variants were novel discoveries due to MI GWAS having less power than single trait GWAS by design. When analyzing the traits with HIPO, 177,519 associated variants were found with 19,829 being new variants. PAT identified 200,112 uniquely associated variants with 22,095 being novel. None of the multi-trait methods identified more distinct variants than the standard single trait GWAS though MTAG, HIPO, and PAT identified new variants. This was likely due to insufficient power to capture variants associated with only one trait.

When comparing the methods on their per trait associations, more associations were identified by leveraging multiple traits. While standard single trait GWAS identified 211,546 uniquely associated variants, only 18,764 were implicated as associated with more than one trait for a total of 233,540 associations as reported in Table 2.3. When analyzing the traits using MTAG, 18,054 out of 164,263 uniquely associated variants were found to be associated with more than one trait. This resulted in there being a total of 184,946 per trait associations. While single trait GWAS and MTAG provided a per trait p-value, MI GWAS, HIPO, and PAT did not. In order to interpret their associations, a per trait m-value must be assigned. When using the m-value framework, MI GWAS interpreted 325,113 per trait associations

due to 96,519 of its 183,669 associated variants being associated with more than one trait. Out of the set of 183,669 uniquely associated variants, there were 8,213 whose interpretation was left ambiguous. This means that while those variants were significantly associated with the set of traits according to the omnibus test, the interpretation as to which of the traits was still ambiguous. HIPO identified 177,519 associated variants where 94,5333 were interpreted as associated with more than one trait. There were 862 with an ambiguous interpretation while 311,509 were interpreted as associated with at least one trait. Finally, the m-value framework was applied to PAT resulting in 363,955 per trait associations from the set of 200,112 unique variants. Out of which 111,126 variants were interpreted as associated with more than one trait and 9,869 were left with an ambiguous interpretation.

We note that while MI GWAS cannot by definition have more power than Single Trait GWAS, once a variant was implicated as associated with at least one trait the interpretation could be assigned to multiple traits. This means that as long as the effect size in one trait was large enough to result in MI GWAS finding the variant significant, the weaker effect sizes could still be interpreted using m-values. This is because the m-value framework leveraged the genetic and environmental covariation between traits regardless of whether or not the original method modeled it which enables an increase in per trait associations. In fact, PAT had over 100,000 more per trait associations than single trait GWAS in Table 2.3 even though it implicated fewer variants. For body mass index, PAT, MI GWAS and HIPO almost doubled the number of per trait associations and nearly tripled it for systolic blood pressure. For diastolic blood pressure, the number of per trait associations was more than tripled due to the m-value framework. In Table 2.3, MTAG performed on par with single trait GWAS on a per trait level. One reason for the difference in performance was the nature of the methods. For MI GWAS, HIPO, and PAT, the variant was first implicated and then interpreted on a per trait basis while MTAG and single trait GWAS assigned statistical significance for each trait separately.

2.2.8 Novel UK Biobank discoveries were replicated in the GIANT consortium

Trait	Power	Average Power		Number of Variants Tested		Expected Number of Replications		Number of Replications		Number of Effect Sizes in Same Direction	
		PAT	HIPO	PAT	HIPO	PAT	HIPO	PAT	HIPO	PAT	HIPO
Body Mass Index	0-10%	6.8%	8.5%	44	5	3	0	0	0	35**	4
	10-20%	15.2%	14.2%	92	24	14	3	0	0	84**	23**
	20-30%	24.7%	25.3%	57	7	14	2	1	0	54**	7*
	30-40%	35.5%	34.2%	40	10	14	3	0	1	37**	9*
	40-50%	45.3%	46.1%	33	12	15	6	2	1	32**	12**
	50-60%	55.5%	55.3%	23	9	13	5	0	1	23**	9**
	60-70%	64.5%	65.4%	43	18	28	12	5	1	40**	18**
	70-80%	75.2%	76.6%	53	66	40	51	1	10	51**	64**
	80-90%	85.9%	82.9%	15	45	13	37	1	8	15**	44**
	90-100%	92.6%	93.6%	8	22	7	21	4	5	7*	19**
Height	0-10%	4.2%	4.9%	15	7	1	0	1	0	12*	7*
	10-20%	13.4%	15.6%	11	30	1	5	2	4	11**	27**
	20-30%	24.7%	24.4%	13	25	3	6	4	2	12**	24**
	30-40%	36.0%	34.6%	30	28	11	10	5	3	26**	27**
	40-50%	45.7%	44.4%	46	18	21	8	5	2	46**	17**
	50-60%	55.5%	55.2%	90	16	50	9	27	0	89**	16**
	60-70%	65.8%	64.6%	135	20	89	13	38	5	130**	20**
	70-80%	75.0%	75.2%	303	21	227	16	92	10	300**	21**
	80-90%	83.1%	83.4%	75	26	62	22	21	14	73**	25**
	90-100%	93.0%	91.7%	17	5	16	5	2	3	17**	5*

Table 2.4: **Replication power in the GIANT consortium for BMI and height.** We tested the novel associations in the UK Biobank discovered by PAT and HIPO for replication in the GIANT consortium. We separately clumped using the lead variant as determined by the m-value. For each variant, we calculate replication power and bin the variants into deciles. The first column lists the trait. The second column is the decile while the third and fourth column are the average power within the set for each respective method. The number of variants tested for replication, the expected number of replications, and the number of variants that replicated are reported in the next six columns. The final two columns contain the number of variants with effect sizes from the GIANT consortium in the same direction seen in the UK Biobank. A binomial test on whether the proportion of effect sizes in the same direction across studies is greater than 50% of all tested variants in the set. A single asterisks means the results are significant at the nominal $\alpha = 0.05$ and two asterisks indicates significance at $\alpha = \frac{0.05}{20}$.

The three methods PAT, HIPO, and MTAG respectively identified 22,095, 19,829, and 931 novel associations when jointly analyzing four traits from the UK Biobank (see Methods). For PAT, all novel associations had an m-value greater than 0.9 in at least one trait which means all associations had a per trait interpretation. The breakdown of the per trait associations were: 12,261 variants interpreted as associated with body mass index, 7,868 with diastolic blood pressure, 21,119 with height, and 7,605 were interpreted as associated with systolic

blood pressure. For HIPO, there were 862 associations with an ambiguous interpretation. The breakdown of the 18,967 variants with a per trait interpretation were: 6,202 associated with body mass index, 8,420 with diastolic blood pressure, 6,396 with height, and 9,844 with systolic blood pressure. For MTAG which provided per trait p-values, 33 of the 931 novel associations were associated with body mass index, 254 with diastolic blood pressure, zero with height, and 644 were associated with systolic blood pressure. Now equipped with novel per trait associations, these discoveries should be validated in an external dataset; therefore, we used the GIANT consortium to see if any of the new associations for body mass index or height could be reproduced [92, 165].

For body mass index, the European summary statistics from the GIANT consortium contained 2,554,638 variants which were separately matched to the variants identified by PAT, HIPO, and MTAG using the RSID, reference, and alternate allele and had a minor allele frequency reported. When the reference and alternate allele differed between the two data sets, the direction of the effect size in the replication data set was flipped. After identifying which variants were present in both data sets, the variants discovered by PAT and HIPO were clumped by taking the largest m-value (i.e., posterior predictive probability) and removing all other variants within a 1MB region. We clumped variants on the m-value instead of the p-value due to the p-value's significance potentially being driven by a different trait. For MTAG, as there was a per trait p-value, we clumped variants using the minimum p-value. Out of the 12,261 novel variants discovered by PAT and interpreted to be associated with body mass index in the UK Biobank, 3,946 were found in the GIANT consortium which resulted in 408 independent variants after clumping. Separately, for the 6,202 variants identified by HIPO, 2,111 were found in the GIANT consortium which resulted in 218 independent variants after clumping. Of the 33 novel variants discovered by MTAG, ten were found in the GIANT consortium which resulted in six independent variants after clumping.

This process was repeated in height where the GIANT consortium had 2,550,859 variants to be considered. Out of the 21,119 variants identified by PAT and interpreted as associated

with height in the UK Biobank, there are 7,068 also found in the GIANT data set. After clumping these variants to the peak m-value per megabase region, there were 735 independent associations. Separately, for the 6,396 variants identified by HIPO, 2,216 were also in the GIANT consortium which resulted in 196 independent variants after clumping. MTAG identified zero novel variants associated with height.

In order to test the replication rate, we performed a one-sided z-test in the direction of the effect size (β) in the UK Biobank. Beginning with PAT, we saw that for body mass index, 378 out of 408 variants (92.6%) had their effect sizes in the same direction in both cohorts. We tested each variant for replication using the level of significance $\alpha = \frac{0.05}{408} = 1.22 \times 10^{-4}$ and found 14 variants replicated. For height, 97.4% (716) of the tested variants had their effect sizes in the same direction. For replication, we set the level of significance to $\alpha = \frac{0.05}{735} = 6.80 \times 10^{-5}$. Here, we saw that 197 of the 735 variants replicated. Separately, we considered the variants discovered by HIPO and saw that for body mass index, 209 of 218 (95.9%) had their effect sizes in the same direction in both cohorts and 27 had a p-value below $\alpha = \frac{0.05}{218} = 2.29 \times 10^{-4}$. For height, 189 of the 196 (96.4%) independent variants discovered by HIPO and interpreted as associated with height had effect sizes in the same direction in both cohorts and 43 had a p-value below $\alpha = \frac{0.05}{196} = 2.55 \times 10^{-4}$. For MTAG, the level of significance for body mass index was $\alpha = \frac{0.05}{6} = 8.33 \times 10^{-3}$. We observed two variants with a p-value below this threshold; all six variants (100%) had their effect sizes in the same direction in both cohorts.

For the variants that failed to replicate, there were a number of possible reasons this occurred. In Table 2.4, we explored how statistical power affected our replication rate. We note that MTAG was not included in the table and that all six variants had a replication power over 90% with a mean of 97.2%. When considering the variants discovered by PAT and HIPO, we first binned the variants into deciles by their replication power. For each decile, we calculated the average power to replicate the effect sizes observed in the UK Biobank in the GIANT consortium. We note that the GIANT consortium did not release the minor allele frequency observed in their samples but instead provided the minor allele frequency

observed in HapMap [70]. While a reasonable estimate, inaccuracies in the minor allele frequency impact power calculations. Additionally, we note that the GIANT consortium summary statistics were from a meta-analysis which may have a lower effective sample size than the reported sample size due to heterogeneity between cohorts. For PAT, the average power over all variants for body mass index was 39.4% while it was 64.1% for HIPO, however we only saw a replication rate of 3.4% and 12.4%, respectively. For height, the replication rate for PAT was 26.8% and was 21.9% for HIPO, but the overall power was 65.5% and 47.3%, respectively. In Table 2.4, we observe for both traits that as power increased so did the replication rate; however, neither trait replicated at the expected rate. The only exception was in height when the power was between 0-30%, we saw PAT replicating slightly over the expected rate.

While we have shown that our replication rate was below expectation, the expected replication rate was likely overestimated due to winner's curse [116]. As the variants tested for replication were not identified as associated by the original single trait GWAS, these variants have sub-optimal power for discovery. This means these variants have small effect sizes and were only found associated after leveraging their covariance structure with other traits. As a result, the bias in the effect size will be much larger here than in variants that were already well powered for discovery. For the non-replicating variants, further power increases (e.g. larger sample sizes) are essential to better tease out which variants warrant follow up analyses.

While many variants failed to replicate potentially due to insufficient power or winner's curse, we also tested whether the effect sizes were in the same direction between GIANT and the UK Biobank. If the variant truly had no effect on the trait, the concordance of effect size across the data sets should be 50%; however, if there was a genetic effect, a higher concordance across data sets is expected. We performed a binomial test in each decile to determine whether the proportion of effect sizes in the same direction was greater than 0.50. Using the significance threshold $\alpha = 0.05$, every test was significant for PAT and 19/20 were significant for HIPO. As there were 20 tests, we adjusted for multiple testing using

a Bonferroni correction ($\alpha = \frac{0.05}{20}$). All but two tests were statistically significant at this new threshold for PAT and 15/20 were significant for HIPO. A test of overall concordance in body mass index tested whether $\frac{378}{408} = 0.926$ was greater than 0.5 returned a p-value of 4.34×10^{-78}. We also tested height ($\frac{716}{735} = 0.974$) which returned a p-value of 1.06×10^{-184}. Separately, for HIPO, we tested the proportion of variants in the same direction in body mass index ($\frac{209}{2018} = 0.959$; $p = 6.43 \times 10^{-51}$) and in height ($\frac{189}{196} = 0.964$; $p = 2.04 \times 10^{-47}$). Therefore, we conclude that while the actual replication rate was low, there is evidence of real genetic signal in the variants identified by the multi-trait methods.

2.2.9 Importance sampling significantly improves PAT's running time

We now show how the cost of null simulations can be reduced using importance sampling. When setting the critical value κ for PAT's likelihood ratio test, the data is simulated according to the null distribution $\mathcal{N}(0, \Sigma_e)$. As a result, a likelihood ratio greater than κ is expected only $\alpha \times n$ times. As GWAS uses the significance threshold of $\alpha = 5 \times 10^{-8}$, the number n needs to extremely large to ensure replication of results, in practice $n = 10^{10}$. Simulating and storing 10^{10} vectors of summary statistics is computationally expensive, especially in terms of memory. This burden can be reduce using importance sampling where the null data is simulated according to a different distribution $\mathcal{N}(0, r\Sigma_e)$ where r is a scaling factor that increases the number of samples that are significant. We note that importance sampling adjusts the weights of the samples in estimating the p-values (see Methods). If r is well chosen κ can be set with fewer simulations.

In Table 2.5, the critical value κ was estimated 25 times and the sample variance of these estimates provided a measure of the stability of the sampling. This was repeated for different values of the scaling factor r and number of samples n. We defined the ratio of the sample variance using importance sampling to the sample variance of null simulations as stability. When the ratio was close to one, the estimated κ using importance sampling was as stable as the κ estimated directly using null simulations and values larger than one indicated importance sampling had a smaller variance. Four traits from the UK Biobank:

Set of Traits (Variance)	Number of Simulations	Scaling of Covariance Matrix ($r \times \Sigma_e$)					
		5	6	7	8	9	10
$\mathcal{B}, \mathcal{D}, \mathcal{H}, \mathcal{S}$ (9e-07)	1e6	6.94	13.82	6.88	12.10	6.49	6.12
	1e5	0.741	1.01	1.40	**1.01**	0.58	0.70
	1e4	0.046	0.12	0.13	0.06	0.09	0.10
$\mathcal{B}, \mathcal{D}, \mathcal{H}$ (5e-07)	1e6	5.27	4.92	7.50	7.11	13.76	7.22
	1e5	0.54	0.53	0.58	**1.15**	1.00	0.69
	1e4	0.035	0.045	0.045	0.074	0.096	0.08
$\mathcal{B}, \mathcal{D}, \mathcal{S}$ (2e-07)	1e6	11.11	7.93	9.35	8.32	11.10	12.07
	1e5	0.62	0.81	1.14	**1.17**	0.90	0.94
	1e4	0.10	0.08	0.16	0.13	0.17	0.18
$\mathcal{B}, \mathcal{H}, \mathcal{S}$ (4e-07)	1e6	3.20	11.75	5.53	6.22	9.29	6.08
	1e5	0.33	0.65	0.64	**0.93**	0.63	0.73
	1e4	0.07	0.06	0.09	0.07	0.09	0.22
$\mathcal{D}, \mathcal{H}, \mathcal{S}$ (9e-07)	1e6	7.23	7.97	18.51	18.31	16.39	12.73
	1e5	0.54	0.73	0.92	**1.30**	1.26	1.47
	1e4	0.10	0.08	0.11	0.08	0.09	0.31
$\mathcal{B}, \mathcal{H}$ (6e-07)	1e6	10.67	12.22	22.48	22.551	20.586	18.295
	1e5	0.59	0.83	0.88	**1.72**	3.02	2.04
	1e4	0.10	0.13	0.20	0.16	0.22	0.27
$\mathcal{D}, \mathcal{S}$ (6e-08)	1e6	28.83	21.53	110.27	27.46	33.39	72.58
	1e5	1.492	4.04	3.49	**4.58**	7.39	5.05
	1e4	0.18	0.37	0.40	0.63	0.51	0.67

Table 2.5: **Stable estimates of critical values in fewer null simulations.** We generate the critical value κ at $\alpha = 5 \times 10^{-8}$ 25 times for various combinations of four traits: body mass index ($\mathcal{B}$), diastolic blood pressure ($\mathcal{D}$), height ($\mathcal{H}$), and systolic blood pressure ($\mathcal{S}$). We simulated data according to $\mathcal{N}(0, r\Sigma_e)$ for $r = \{5, 6, 7, 8\}$ and for $n = 10^4$, 10^5 and 10^6 simulations. We then take a ratio of the variation in the estimated critical value κ which we call the stability. The first column is the set of traits and the variance for $\mathcal{N}(0, 1\Sigma_e)$ using $n = 10^{10}$ simulations. The second column is the number of simulations while the remaining columns show the stability for different scaling factors of the covariance matrix $r : r = \{5, 6, 7, 8\}$.

body mass index ($\mathcal{B}$), diastolic blood pressure ($\mathcal{D}$), height ($\mathcal{H}$), and systolic blood pressure ($\mathcal{S}$) were considered in these simulations as well as subsets of the traits [110, 147]. When using 10^6 simulations, we found using importance sampling was consistently more stable for all reported scaling factors, r. For diastolic blood pressure and systolic blood pressure, importance sampling was also more stable for all reported scaling factors using 10^5 simulations.

When using the scaling factor $r = 8$ for $n = 10^5$ simulations, the variance for the value κ when using importance sampling was approximately equal to the variance using 10^{10} null simulations across the various sets of traits. For most sets of traits importance sampling was still slightly more stable; it was only for body mass index, height, and systolic blood pressure that it was less stable with a ratio of 0.93 which is still very close to 1. This means, the same stability could be achieved using only 10^5 simulations which is 10^5 fewer simulations. This reduction in computational resources holds true across data sets. In practice, however, the use of 10^6 simulations is more practical as the stability of the critical value κ is less sensitive to the setting of r. This still results in using $10,000$ fewer simulations.

2.3 Discussion

Here, we presented PAT, a method that leveraged pleiotropy for joint association testing in multiple traits as well as an extension to the m-value framework. Through simulations, PAT was shown to control the false positive rate as well as significantly increase statistical power to detect pleiotropic effects. The impact of misspecifying model parameters on PAT was also explored. We saw that PAT was robust to there being a genetic effect in some subsets of traits while other configurations significantly impacted PAT's performance. One major limitation of PAT was its lack of per trait interpretations. This was overcome by the extension to the m-value framework presented here. M-values enabled a per trait interpretation of PAT and other omnibus methods. Through simulations, we found that the false positive assignment rate from m-values was low.

Additionally, PAT was compared to three multi-trait methods: MTAG, HIPO, and AS-SET. While PAT was shown to be a more powerful method for omnibus association testing, there were some scenarios where PAT was underpowered. One such scenario was when there was high environmental correlation. In this scenario, HIPO and MTAG provided better models for joint analysis of traits due to PAT's conservative nature in the presence of strong environmental correlation.

While we primarily considered how the methods handled the misspecification of the covariance structure between traits. Another fundamental difference between PAT and the other methods was how PAT derived its critical value using null simulations. In contrast, the other methods produced their p-values analytically; however, they could also leverage null simulations. This may be particularly beneficial to HIPO whose signal was likely spread across multiple components. Accounting for this empirically may better calibrate the global null. Further work would need to be done to explore this, but we note that importance sampling as used here would enable an efficient solution.

In addition to simulations, PAT analyzed four traits in the UK Biobank and discovered 22,095 novel associations while the next best method, HIPO, identified 19,829. After computing m-values and clumping the per trait associations, the replication of associated variants in body mass index and height were tested in the GIANT consortium. For body mass index, 14 of 408 independent variants discovered by PAT replicated while 27 of 218 independent variants replicated for HIPO. The replication rate in height was much higher with 197 out of 735 variants replicating for PAT and 43 of 196 for HIPO. While the replication rate was below expectation, this may be due to winner's curse[116]. The variants tested for replication were novel discoveries that were under powered in the original association. In addition to testing replication, we tested whether the effect sizes between the UK Biobank and GIANT consortium were in the same direction. Overall, there was significant evidence that the direction of the effect sizes were concordant which is improbable under the null.

While PAT was shown to be an effective method for leveraging pleiotropy between traits, the optimal number of traits to jointly model was not explored. As the number of traits increase, the genome-wide estimate of genetic correlation ceases to hold across all traits. This would result in fewer novel associations as the power gains would be stunted by model misspecification. Further exploration is needed to determine which traits should be analyzed together and how to effectively cluster the traits into these sets.

Another limitation to PAT was it assumed the genetic covariance structure was constant across the genome. PAT was agnostic to the environmental and genetic covariance between

traits and treated these as input. As all variants were tested independently, the user could input a different covariance structure for each variant or a set of variants. This may enable a significant power increase as modeling local covariance structure better reflects the covariance structure between z-scores [46, 139, 140]. Our method, however, only considered the global estimate of genetic and environmental correlation between traits and further work is needed to quantify the impact of such modifications on both power and false positives which other's have explored [84].

One limitation to the m-value interpretation framework was how it estimated the number of causal variants for the genetic covariance matrix. Currently, for association testing the genetic covariance matrix was scaled according to the polygenic model (i.e. all variants were causal). Once variants were implicated as associated, we used grid search to find the genetic covariance matrix scaling that best reflected the average effect size of independent variants. This in effect was an approximation to the number of causal variants. Further work is merited to better estimate the number of causal variants for each trait as well as the number shared between traits.

2.4 Materials and Methods

2.4.1 Association testing in a single quantitative trait (GWAS)

We now describe the standard approach for determining if a genetic variant g is associated with a quantitative trait y. Let y and g be measured for N individuals where $g_j \in \{0, 1, 2\}$ is the minor allele count for each individual j. The column vector g is then standardize according to the population proportion of the minor allele p where $2p$ is the mean and $2p(1-p)$ is the variance of g. This standardized column vector x is defined as follows $x : x \in \left\{ \frac{-2p}{\sqrt{2p(1-p)}}, \frac{1-2p}{\sqrt{2p(1-p)}}, \frac{2-2p}{\sqrt{2p(1-p)}} \right\}$. The quantitative trait y is normally distributed such that $y \sim \mathcal{N}(\mu, \sigma_e^2 I)$ where μ is the mean and σ_e^2 is the variance of the trait. This can then be mean-centered and scaled which results in the column vector $y \sim \mathcal{N}(0, I)$. We can now assume the following linear model:

$$y = \beta x + e \tag{2.2}$$

where β is the effect size of the variant x on the trait y and the error e follow the standard normal [42]. Ordinary Least Squares results in the estimator $\hat{\beta} = \frac{x^T y}{N}$ where $\hat{\beta} \sim \mathcal{N}(\beta, \frac{1}{N})$. Setting $s = \frac{\hat{\beta}}{\hat{\sigma_e}}\sqrt{N}$ results in the following Gaussian: $s \sim \mathcal{N}(\frac{\hat{\beta}\sqrt{N}}{\hat{\sigma_e}}, 1)$.

We now test the null hypothesis: x is not associated with y. More formally this tests if $\beta = 0$ or $s \sim \mathcal{N}(0, 1)$. The null model is rejected if $|s| > z$ where z is the z-statistic at the α level of significance for the standard normal distribution. The corresponding critical value $z = \Phi^{-1}(1 - \frac{\alpha}{2})$. Typically, human GWAS uses $\alpha = 5 \times 10^{-8}$ [70, 104, 119].

2.4.2 Generalizing GWAS testing to multiple traits (MI GWAS)

As previously stated, GWAS traditionally analyzes each trait y_i in a set of T traits independently. In fact, each trait may be measured on distinct sets of individuals. Let us assume none of the traits $y_1, .., y_T$ have overlapping individuals; therefore every trait y_i and the standardized genetic variant x is measured for N_i individuals. This assumption will later be relaxed. For now, the z-score for trait i: s_i is tested for whether $|s_i| > z$, and this process is repeat for each trait independently. Another approach instead of performing T different hypothesis tests is to determine whether the variant is associated with at least one of the traits. The corresponding null hypothesis is the variant is not associated with any of the traits. We refer to this method as multiple independent GWAS (MI GWAS). This results in $\beta_1 = \ldots = \beta_T = 0$ which is equivalent to saying the null model is $s_1 \sim \mathcal{N}(0, 1), \ldots, s_T \sim \mathcal{N}(0, 1)$. A simple way to test our null hypothesis is to check if the largest $s_i \in S = \{|s_1|, \ldots, |s_T|\}$ is greater than the critical value z though z will now need to be corrected for multiple testing. This can be done using a Bonferroni correction for the number of traits, T, so the critical value is $z = \Phi^{-1}(1 - \frac{\alpha}{2T})$.

Another method for setting the critical value is using null simulations. This is done by simulating data according to $S = \{s_1, \ldots, s_T\}$ such that every s_i is under the null hypothesis. As all traits are measured for different groups of individuals, there is no covariation between any pairs of traits. This means the multivariate $S \sim \mathcal{N}(0, \Sigma_e)$ has the identity matrix as its covariance matrix; therefore, we simulate $S \sim \mathcal{N}(0, I)$ n times keeping the $\max\{|s_1|, \ldots, |s_T|\}$ for each S. We then sort the n retained values and assign a p-value to each critical value using the quantile.

2.4.3 Using pleiotropy for association testing in multiple traits (PAT)

Another method for hypothesis testing is a likelihood ratio test which compares the null model to a proposed alternative model. Currently, only the null model has been defined. For a single quantitative trait y whose null hypothesis is $\beta = 0$ and $s \sim \mathcal{N}(0, 1)$, the alternative hypothesis is $\beta \neq 0$ and β is assumed to follow a Gaussian distribution: $\beta \sim \mathcal{N}(0, \sigma_g^2)$, where σ_g^2 is the additive, per-variant heritabilty of the trait. As Gaussian distributions are conjugate priors to Gaussian likelihood functions, the distribution of β can be used to get the Gaussian posterior predictive distribution, $s \sim \mathcal{N}(0, 1 + N\sigma_g^2)$. Two models that describe s have been defined and result in the following likelihood ratio:

$$\frac{P(s|\mu = 0, \sigma^2 = 1 + N\sigma_g^2)}{P(s|\mu = 0, \sigma^2 = 1)} > \kappa \tag{2.3}$$

If the ratio of the likelihood functions is larger than κ, the null hypothesis is rejected. Before expounding on how to set κ, we will first extend the likelihood ratio test to the case of multiple traits. We retain the assumption that the traits are not measured on the same individuals. This means, there is no environmental correlation, so under the null hypothesis $S \sim \mathcal{N}(0, I)$. The assumption about distinct sets of individuals does not, however, have the same implication for the genetic correlation between genetic effects. Letting the $\text{cov}(\beta_i, \beta_k) = \sigma_{g_{i,k}}$ we can derive $\text{cov}(s_i, s_k) = \sqrt{N_i}\sqrt{N_k}\sigma_{g_{i,k}}$. This results in the alternative model being:

$$S \sim \mathcal{N}\left(\begin{pmatrix}0\\ \vdots \\ 0\end{pmatrix}, \begin{pmatrix}1+N_1\sigma^2_{g_1} & \dots & \sqrt{N_1}\sqrt{N_T}\sigma_{g_{1,T}} \\ \vdots & \ddots & \vdots \\ \sqrt{N_1}\sqrt{N_T}\sigma_{g_{1,T}} & \dots & 1+N_T\sigma^2_{g_T}\end{pmatrix}\right) \tag{2.4}$$

which can be written as $S \sim \mathcal{N}(0, \Sigma_e + \Sigma_g)$. This means under the alternative model, the covariance of S is the sum of the environmental and genetic covariance where for now the environmental covariance is still the identity matrix, I. The likelihood ratio for PAT is now defined as:

$$\frac{P(S|\mu = 0, \Sigma = \Sigma_e + \Sigma_g))}{P(S|\mu = 0, \Sigma = \Sigma_e)} = \frac{P(S|\mu = 0, \Sigma = I + \Sigma_g))}{P(S|\mu = 0, \Sigma = I)} > \kappa \tag{2.5}$$

The critical value κ is set for PAT using the same null simulations of $S \sim \mathcal{N}(0, I)$. This time the likelihood ratio for each S is retained, sorted, and assigned a p-value using the quantile.

2.4.4 Modeling overlapping samples across traits

We now relax the assumption that no individual is measured for more than one trait. Under the null hypothesis, this means that Σ_e in $S \sim \mathcal{N}(0, \Sigma_e)$ would not be the identity matrix I. In this case, $cov(s_i, s_k) = \frac{N_{shared}}{\sqrt{N_i}\sqrt{N_k}}\rho_{e_{i,k}}$. This means the covariance between s_i and s_k is the environmental correlation between the traits, and the environmental correlation is weighted by the proportion of overlapping individuals. Under the alternative hypothesis, we have $S : S \sim \mathcal{N}(0, \Sigma_e + \Sigma_g)$. We note that while sample overlap between traits affects Σ_e, it does not impact Σ_g.

2.4.5 Leveraging importance sampling for null simulations

When performing null simulations, the number of simulations n must be large enough that the critical value κ is stable across estimates. In practice, this can require n to be very large when α is really small because simulating $S : S \sim \mathcal{N}(0, \Sigma_e)$ with a likelihood ratio larger

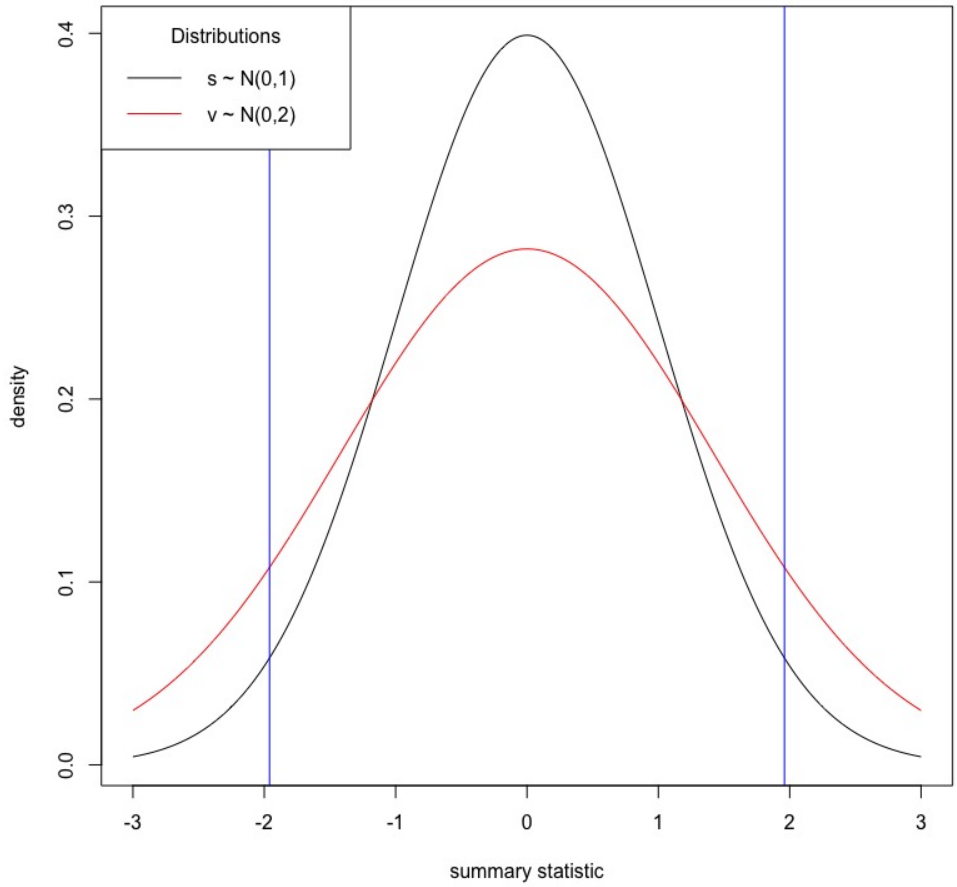


Figure 2.5: **Using importance sampling for setting critical values.** We simulated data according to two univariate Gaussian distributions $s \sim \mathcal{N}(0,1)$ and $v \sim \mathcal{N}(0,2)$ and show the densities. We show the critical value $z \approx 1.96$ for $\alpha = 0.05$. We would expect to see the critical value $|z|$ or larger more often when simulating data according to v than when simulated under the distribution of s.

than κ is expected to occur $\alpha \times n$ times. One method for reducing the number of simulations is importance sampling.

To explain our approach we first review importance sampling in one trait. While, traditionally $z = \Phi^{-1}(1 - \frac{\alpha}{2})$ is used to set the critical value z for the standard normal. It is also possible to use null simulations just as we do for MI GWAS and PAT. We simulate $s \sim \mathcal{N}(0,1)$ n times and sort $|s|$ and assign the p-values using the quantile. To obtain the significance of a specific critical value such as 5.2, enough null simulations must be performed

to have a sufficient number of samples above the critical value. The p-value would then be estimated by counting the number of samples above the critical value divided by the total number of samples. Unfortunately, for very significant p-values this requires a very large number of samples since the vast majority of samples are below the critical value.

Importance sampling reduces the number of simulations needed for setting the critical value by simulating data according to a different distribution v where v results in samples larger than the critical value z to occur more frequently. The procedure for estimating the p-value will then be adjusted to account for the differences between the two distributions, s and v. In our approach, v has the following distribution $v \sim \mathcal{N}(0, r1)$, and in Fig 2.5 the scaling factor $r = 2$ is used. In this figure, the critical value $z \approx 1.96$ for $\alpha = .05$ is shown for the null distribution s. We can see in Fig 2.5 that the distribution v has many more samples in the tails; therefore, the significance level α does not correspond to z for the distribution v. The p-value using importance sampling is estimated for each data point by first computing a weight w. This weight w is the likelihood ratio $\frac{P(v|\mu=0,\sigma=1)}{P(v|\mu=0,\sigma=2)}$ of the data points from v under the two models. By summing the weights of samples larger than the critical value and dividing by the sum of the weights for all samples, the p-value can be set for each critical value. We note that if $r = 1$, then s and v are identical and all the weights are 1. In this case, importance sampling and the standard approach are equivalent.

We can now extend this to learning about the null distribution of $S \sim \mathcal{N}(0, \Sigma_e)$ by simulating data according to the distribution $V : V \sim \mathcal{N}(0, r\Sigma_e)$. Again we will find that a well chosen alternative distribution V results in more statistics greater than κ in fewer simulations. The weight w is the likelihood ratio $\frac{P(V|\mu=0,\Sigma=\Sigma_e)}{P(V|\mu=0,\Sigma=r\Sigma_e)}$ and will be used to obtain p-values as described above. When picking the scaling factor r, larger values will sample the tail in fewer simulations, however, care should be taken to ensure that a sufficient number of simulations are used to accurately set the critical threshold κ. In Table 2.5, we found that 10^6 consistently provided a stable estimate of the critical value for $\alpha = 5 \times 10^{-8}$ regardless of the choice of r.

2.4.6 Interpreting GWAS omnibus associations

When performing an omnibus hypothesis test, there is only one p-value which cannot be directly interpreted on a per trait level. In previous work, the statistic m-values were introduced to enable interpretation of GWAS meta-analyses across studies, with m-values being the posterior probability of a genetic effect per study [63] . The original m-values assumed that across studies the effect sizes are similar as it considered the same trait across multiple studies. When applying this framework to multiple traits, the model needs to account for differing effect sizes to prevent spurious results. Below, we describe the extension to the m-value framework which assumes a random effects model for the genetic effect and that the effect sizes reflect the genome-wide estimate of genetic correlation.

We assume there are T traits for which a variant has been identified as associated by PAT (or another omnibus test). While the variant is known to be associated, there are many possible configurations of an effect. There may be an effect in all traits in which case the configuration is $c = (1, ..., 1)$, or there may only be an effect in the first trait, $c = (1, 0, ..., 0)$. The set of all configurations can be written as $C = \{0, 1\}^T$ where $|C| = 2^T$. For each trait i, there is subset of configurations $C_i \subset C$ that are compatible with the variant having a genetic effect in that particular trait, where $|C_i| = 2^{T-1}$. This means that for every configuration $c \in C_i$, the ith index is always 1.

When PAT determines a variant is associated with the set of T traits, it assumes the variant affecting all T traits; therefore, the assumed posterior predictive distribution of effect is $P(S|\mu = 0, \Sigma = \Sigma_g + \Sigma_e)$. While pleiotropy is ubiquitous, the assumption that every variants affects all traits is not realistic. We will now define the genetic covariance matrix $\Sigma_g(c)$ that corresponds to a configuration c where

$$\Sigma_g(c) = \begin{cases} \sqrt{N_i}\sqrt{N_j}\sigma_{g_{i,j}}, & \text{if } \ c_i = 1 \text{ and } c_j = 1 \\ 0, & \text{otherwise} \end{cases} \tag{2.6}$$

We note that when $c = (1, ..., 1)$, $\Sigma_g = \Sigma_g(c)$. With this in mind, m-values works by summing

the posterior probabilities that corresponding to the configurations in C_i and dividing by the the total sum of all posterior probabilities (set of configuration in C). Therefore, for each trait i:

$$m_i = \frac{\sum_{c \in C_i} P(S|\mu = 0, \Sigma = \Sigma_e + \Sigma_g(c))}{\sum_{c \in C} P(S|\mu = 0, \Sigma = \Sigma_e + \Sigma_g(c))} \tag{2.7}$$

where S are summary statistics across the T traits for one variant, and the m-value m_i is the proportion of the all posterior probabilities compatible with there being an effect in trait i. When $m_i > 0.9$, the variant is assumed to be associated with the ith trait. Otherwise, the interpretation is left ambiguous.

While we assume the covariance structure of Σ_g follows the polygenic model, for interpretation purposes this assumption is relaxed. Under the polygenic model, every variant has an effect; therefore, the expected effect size of each variant is $\frac{1}{M} \times h^2$ where M is the total number of variants and h^2 is the estimated additive heritability of the trait. When only considering the variants found genome-wide significant, the expected effect size of these variants needs to be to estimated. We do this by estimating the number of causal variants Q and rescale the genetic covariance matrix Σ_g by $\frac{M}{Q}$ for the m-value interpretation framework.

This is necessary because $h^2 \in [0, 1]$ and with genome-wide association studies using millions of variants, $\Sigma_g + \Sigma_e \approx \Sigma_e$ under the polygenic model. While a valid model for association testing, distinguishing between different configurations of $\Sigma_g(c)$ to calculate the m-value is very difficult. Therefore, we scale Σ_g and the resulting $\Sigma_g(c)$ by randomly selecting one associated variant per 100KB region for a total of k variants. We then perform a grid search for $Q \in [1, M]$ and retain the value of Q which maximizes the likelihood function as shown below:

$$\underset{Q \in [1,M]}{\operatorname{argmax}} \prod_{i=1}^{k} P(S_i|\mu = 0, \Sigma = \frac{M}{Q}\Sigma_g + \Sigma_e) \tag{2.8}$$

Trait	Genetic Variance	Sample Size	LD-Score Intercept
body mass index ($\mathcal{B}$)	0.241	336,107	1.082
diastolic blood pressure ($\mathcal{D}$)	0.129	317,756	1.074
height ($\mathcal{H}$)	0.429	336,574	1.237
systolic blood pressure ($\mathcal{S}$)	0.114	317,754	1.108

Table 2.6: **The genetic variance and sample sizes from the 2017 UK Biobank release.** We used summary statistics for body mass index, diastolic blood pressure, height, and systolic blood pressure from the 2017 release of the UK Biobank as input for simulations. We reported the sample sizes, genetic variance estimated by LD-Score regression, and the LD-Score intercept.

Trait 1	Trait 2	Genetic Correlation	Environmental Correlation	LD-Score Intercept
body mass index ($\mathcal{B}$)	diastolic blood pressure ($\mathcal{D}$)	0.305	0.258	0.303
body mass index ($\mathcal{B}$)	height ($\mathcal{H}$)	-0.165	-0.084	-0.137
body mass index ($\mathcal{B}$)	systolic blood pressure ($\mathcal{S}$)	0.166	0.198	0.219
diastolic blood pressure ($\mathcal{D}$)	height ($\mathcal{H}$)	-0.125	-0.004	-0.025
diastolic blood pressure ($\mathcal{D}$)	systolic blood pressure ($\mathcal{S}$)	0.648	0.686	0.765
height ($\mathcal{H}$)	systolic blood pressure ($\mathcal{S}$)	-0.144	-0.104	-0.132

Table 2.7: **The genetic and environmental correlation used from the 2017 version of UK Biobank data.** We used summary statistics for body mass index, diastolic blood pressure, height, and systolic blood pressure from the 2017 release of the UK Biobank as input to simulations. We used cross-trait LD-Score regression to estimate the genetic and environmental correlation and report the LD-Score intercept.

Trait	Genetic Variance	Sample Size	LD-Score Intercept
body mass index ($\mathcal{B}$)	0.243	359,983	1.057
diastolic blood pressure ($\mathcal{D}$)	0.132	340,162	1.068
height ($\mathcal{H}$)	0.469	360,388	1.270
systolic blood pressure ($\mathcal{S}$)	0.139	340,159	1.070

Table 2.8: **The genetic variance and sample sizes used in simulations and real data analyses.** Using the summary statistics for body mass index, diastolic blood pressure, height, and systolic blood pressure from the 2018 release of the UK Biobank, we estimated the genetic variance with LD-Score regression. We report the sample sizes, genetic variance, and LD-Score intercept.

2.4.7 Description of the UK Biobank data

We used four traits from the UK Biobank released in 2017 and in 2018 as the basis of our simulations [110, 111]. For both sets of summary statistics, only the variants which were

Trait 1	Trait 2	Genetic Correlation	Environmental Correlation	LD-Score Intercept
body mass index ($\mathcal{B}$)	diastolic blood pressure ($\mathcal{D}$)	0.305	0.228	0.275
body mass index ($\mathcal{B}$)	height ($\mathcal{H}$)	-0.164	-0.066	-0.121
body mass index ($\mathcal{B}$)	systolic blood pressure ($\mathcal{S}$)	0.174	0.15	0.177
diastolic blood pressure ($\mathcal{D}$)	height ($\mathcal{H}$)	-0.122	-0.001	-0.030
diastolic blood pressure ($\mathcal{D}$)	systolic blood pressure ($\mathcal{S}$)	0.663	0.678	0.768
height ($\mathcal{H}$)	systolic blood pressure ($\mathcal{S}$)	-0.151	-0.047	-0.083

Table 2.9: **The genetic and environmental correlation used in real data analyses and simulations.** We used summary statistics for body mass index, diastolic blood pressure, height, and systolic blood pressure from the 2018 release of the UK Biobank as input to cross-trait LD-Score regression. We report the genetic and environmental correlation as well as the LD-Score intercept.

biallelic, have non-ambiguous strands, a minor allele frequency greater than 1%, an INFO score greater than 80%, and found in the 1000 Genomes European reference panel were retained [1]. We used LD-Score regression [18] to calculate the genetic variance of each trait as shown in Table 2.6 and Table 2.8. For calculating genetic covariance and environmental covariance, we used cross-trait LD-Score regression [17]. The genetic covariance produced by the software was reported in Table 2.7 and Table 2.9. By taking the intercept and scaling it by $\frac{\sqrt{N_1 N_2}}{N_s}$ where N_1 is the sample size in trait 1, N_2 is the sample size in trait 2, and N_s is the number of overlapping individuals, we were able to recover the phenotypic covariance. By subtracting the genetic covariance from the phenotypic covariance, the environmental covariance is estimated. As we used summary statistics and the true sample overlap was unknown, we assumed there were no trait specific missing individuals, and we, therefore, set $N_s = \min\{N_1, N_2\}$.

For simulations, PAT used the reported values to define its likelihood ratio test; MI GWAS, however makes no assumptions about the phenotypic covariance. HIPO and MTAG defined their parameters slightly differently, but all methods used the same results from LD-Score regression and cross-trait LD-Score regression. Simulations showing the stability of null simulations used the 2017 version of the UK Biobank summary statistics (see Table 2.6 and Table 2.7 for values). All other simulations and real data analyses were based on the summary statistics from 2018 (see Table 2.8 and Table 2.9). The switch in summary statistic

version was due to the 2017 version of the UK Biobank results being no longer available which prevented reproduction of results.

CHAPTER 3

Identifying causal variants by fine mapping across multiple studies

3.1 Introduction

Genome-wide association studies (GWAS) have successfully identified numerous genetic variants associated with a variety of complex traits in humans [48, 69, 92]. However, most of these associated variants are not causal, and are simply in linkage disequilibrium (LD) with the true causal variants. Identifying these causal variants is a crucial step towards understanding the genetic architecture of complex traits, but testing all associated variants at each locus using functional studies is cost-prohibitive. This problem is addressed by a statistical approach known as fine mapping, which attempts to prioritize a small subset of variants for further testing while accounting for their correlation structure [134].

The classic approach to fine mapping involves simply selecting a given number of SNPs with the strongest association statistics for follow-up, but this performs sub-optimally because it does not account for LD [45]. Bayesian methods that did account for LD were developed, but were based upon the simplifying assumption that each locus only harbors a single causal variant [11, 98]. This assumption, however, is not true in many cases [68]. Additionally, many early methods required individual-level genetic data, whereas many human GWAS often provide only summary statistics due to privacy concerns. CAVIAR introduced a Bayesian approach that relied only on summary statistics and LD, accounted for uncertainty in association statistics using a multivariate normal (MVN) distribution and allowed for the possibility of multiple causal SNPs in a locus [68]. This approach was widely adopted and

later made more efficient by methods such as CAVIARBF, FINEMAP and JAM [13, 25, 112].

There is growing interest in improving fine mapping by leveraging information from multiple studies. One of the most important examples of this is trans-ethnic fine mapping, which can significantly improve fine mapping power and resolution by leveraging the distinct LD structures in each population [36, 120, 163]. The benefits of which have been shown through methods such as trans-ethnic PAINTOR and MR-MEGA [75, 97]. Intuitively, the set of SNPs that are tightly correlated with the causal SNP(s) will be different in different populations, allowing more SNPs to be filtered out as potential candidates. However, the varying LD patterns also present a unique challenge in the multiple study setting that trans-ethnic fine mapping methods must handle. Additionally, while there is evidence that the same SNPs drive association signals across populations, there is also heterogeneity in their effect sizes which presents another challenge [83, 83, 101, 101, 120, 163]. Existing methods either assume a single causal SNP in each locus or do not explicitly model heterogeneity which limits thier power [62, 75, 97, 109].

In this chapter, we present MsCAVIAR, a novel method that addresses these challenges. We retain the Bayesian MVN framework of CAVIAR while introducing a novel approach to explicitly account for the heterogeneity of effect sizes between studies. Our method requires only summary statistics and LD matrices as input, allows for multiple causal variants in a locus, and models uncertainty in association statistics and between-study heterogeneity. The output is a set of SNPs that contains all causal SNPs in the locus, at a user-set confidence threshold (e.g. 95%). We show in simulation studies that MsCAVIAR outperforms existing trans-ethnic fine mapping methods and extensions of CAVIAR to the multiple study setting [68, 75]. We further demonstrate the efficacy of MsCAVIAR in a real data application.

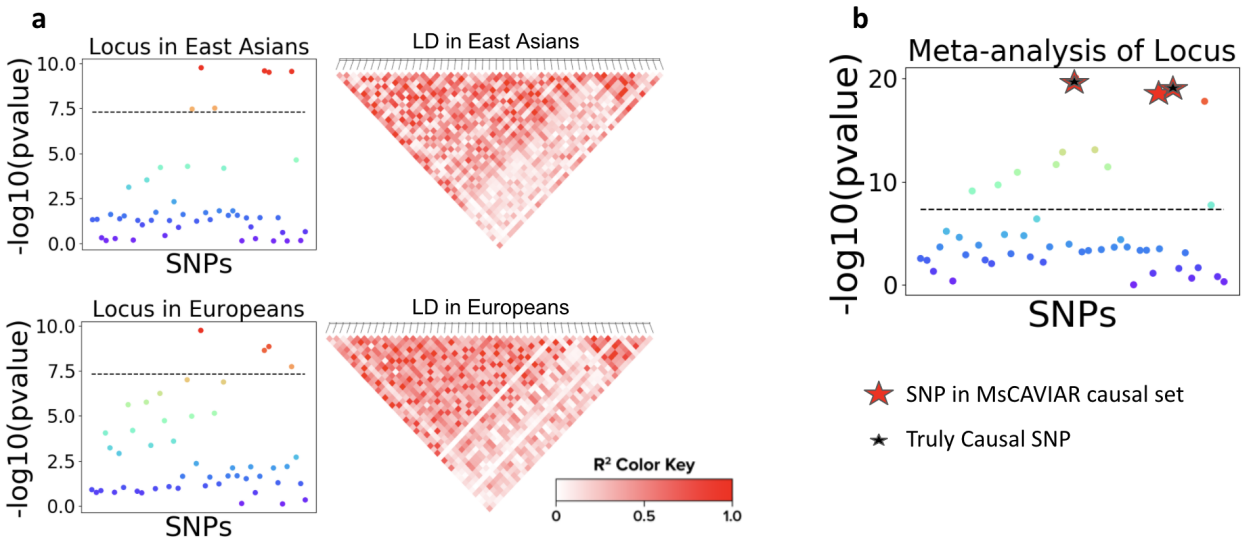


Figure 3.1: **Overview of MsCAVIAR.** (A) Simulated z-scores for SNPs at one locus in two different ancestral populations: East Asian (top) and European (bottom), shown by their $-log_{10}$(p-value). LD matrices for these populations were derived using data from the 1000 Genomes project and treated as input for MsCAVIAR. (B) Meta-analysis results for this locus, showing many significant SNPs. Also displayed are the SNPs that are in the causal set that MsCAVIAR returns (red stars) and the truly causal SNPs (black stars).

3.2 Results

3.2.1 Methods overview

Our method, MsCAVIAR, takes as input the association statistics (e.g. z-scores) and linkage disequilibrium (LD) matrix for SNPs in one locus from each study (Fig 3.1 sub-figure A). The LD matrix can be computed from in-sample genotyped data or appropriate reference panels such as the 1000 Genomes projector HapMap project [1, 70]. MsCAVIAR computes and outputs a minimal-sized set of SNPs that, with probability at least ρ, contains *all* causal SNPs known as the causal set; ideally this causal set contains far fewer SNPs than the set of significant SNPs obtained via a direct meta-analysis (Fig 3.1 sub-figure B).

By our definition of a causal set, every causal SNP must be contained in the set with high probability, but not every SNP in the set needs to be causal. Concretely, each SNP can be assigned a binary causal status: 1 for causal or 0 for non-causal. So long as none of the SNPs outside of the causal set are set to 1, the assignments are compatible with our definition of

a causal set. We can represent these causal status assignments in a binary vector with one entry for each SNP denoting its causal status; we call such a vector a configuration and denote it as C. For each configuration C compatible with the causal set, we compute its posterior probability in a Bayesian manner: the probability of a configuration of SNPs being causal given the association statistics can be computed by modeling a prior probability for that configuration and a likelihood function for the association statistics given the assumed causal SNPs given by C (see Methods).

The overall likelihood function can be decomposed into a product over the likelihood function for each study, since we assume that the studies are independent. More specifically, we assume that there is a true global effect size for a SNP over all possible populations, around which the effect sizes for that SNP in different studies are independently drawn according to a heterogeneity variance parameter. This allows MsCAVIAR to model the fact that effect sizes of a SNP across different studies are related, but not equal. Because we expect the summary statistics to be a function of their LD with the causal SNPs, the parameters of the likelihood function for each study are different, assuming the studies have different LD patterns. By computing the product over the likelihood of each study, we are able to account for their different LD patterns to determine the likelihood over all studies.

The posterior probability for a causal set is then computed by summing the posterior probabilities of all compatible configurations, and then dividing by the sum of the posterior probabilities for all possible configurations. We start by assessing causal sets containing only one SNP, and then causal sets containing two SNPs, and then three SNPs, and so on until a causal set exceeds the posterior probability threshold ρ. In practice, ρ is set to a high value such as 95%.

3.2.2 MsCAVIAR improves fine mapping resolution in a simulation study

We now describe our simulation study to evaluate the performance of MsCAVIAR as compared with other methods. We selected two samples of 9,000 unrelated individuals from the UK Biobank, one with European ancestry and the other with Asian ancestry [147]. In order

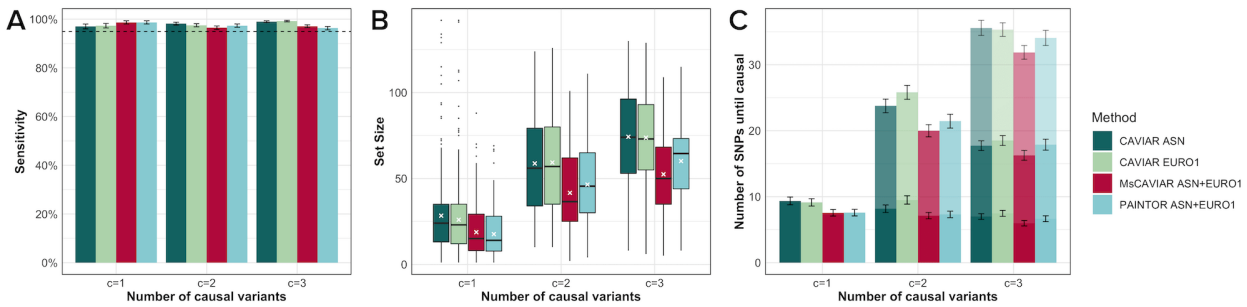


Figure 3.2: **Comparison of sensitivity, precision, and set sizes using simulated data.** We compare MsCAVIAR, PAINTOR, and CAVIAR with $c \in \{1, 2, 3\}$ causal variants implanted with results averaged over 20 replicates for 3 loci and 5 levels of heritability for all 3 values of c. (A) Bar graph indicating the sensitivity of each method with a dashed line to reflect the expected posterior probability, $\rho = 0.95$, of recovering all causal SNPs. (B) Box plots showing the average set sizes returned by the methods. Each box is the interquartile range of causal set sizes with the middle black line representing the median, and the white crosses showing the mean. (C) Bar graph displaying the average the number of SNPs in descending order of posterior inclusion probability (PIP) until 1, 2, or 3 causal SNPs are identified. Stacked bars represent increasing numbers of causal SNPs identified, until the true number of causal SNPs (x-axis) are identified.

to generate realistic fine mapping scenarios, we centered 100KB windows around SNPs that reached genome-wide significant association with high-density lipoprotein (HDL) cholesterol in the UK Biobank summary statistics released by the Neale lab [111]. From these windows, we selected three loci that reflected high, medium, and low patterns of LD as defined by the proportion of SNPs with at least 90% LD (32%, 25%, and 8%, respectively). We then obtained the imputed genotype data for these loci for our samples in the UK Biobank. The loci were filtered for missing genotypes ($> 0\%$) and low minor allele frequency ($< 1\%$). The loci with low, medium, and high LD had 144, 126, and 154 SNPs, respectively.

We then simulated causal SNPs and their effect sizes $\beta \sim \mathcal{N}(\frac{5.2}{\sqrt{9000}}, 1)$, independently for the cases of 1, 2, or 3 causal SNPs randomly chosen within each locus. For simplicity, we take the absolute value of the effect size and restrict causal SNPs to being positively correlated with each other. We then used GCTA to simulate phenotypes using different heritability levels: 0.2%, 0.4%, 0.6%, 0.8%, and 1%, times the number of causal SNPs [168]. Concretely, GCTA simulates the phenotype y according to $y = X\beta + e$, where X is the standardized genotype matrix for the causal variant(s), β is the vector of causal variant

effect sizes, and e is a vector of environmental noise terms where $e = \sigma_g^2(1/h^2 - 1)$. In other words, the environmental variance is scaled to achieve the desired heritability. Thus, modulating the heritability affects the strength of the association signal between variants and the phenotype, while drawing different effect sizes for different causal variants allows for the modeling of heterogeneity. Finally, we run a linear regression using fastGWA to generate the summary statistics [71]. We simulated 20 replicates (re-drawing the causal SNPs and their effect sizes) for each level of heritability and number of causal SNPs for a total of 900 simulations.

Using this data, we compared MsCAVIAR to the trans-ethnic mode of PAINTOR and to CAVIAR run on Asians and Europeans, individually (Fig 3.2) [68, 75]. For each number of causal SNPs (1, 2, or 3), we averaged the results across all simulated scenarios. For each method, we provided the in-sample LD and the summary statistics described above. All methods were run with posterior probability threshold $\rho^* = 0.95$, so methods with 95% or higher sensitivity were considered well-calibrated (dashed line in Fig 3.2 sub-figure A). MsCAVIAR's heterogeneity parameter was set to $\tau^2 = 0.52$ (see Methods). We also evaluated methods for the size of their returned causal sets (Fig 3.2 sub-figure B) because, conditioned on having a well-calibrated recall, it is preferable to return a small causal set. This can be thought of as higher precision, as non-causal SNPs in the causal sets can be thought of as false positives.

All of the methods in this assessment were well-calibrated (Fig 3.2 sub-figure A), which is expected, as previously shown for CAVIAR and PAINTOR. For each number of causal SNPs, MsCAVIAR and PAINTOR returned substantially smaller set sizes than CAVIAR run on either population individually, highlighting the benefit of utilizing information from multiple studies. With one causal SNP in the locus, MsCAVIAR and PAINTOR had similar causal set sizes, with MsCAVIAR's mean and median set sizes being 18.7 and 15.0 and PAINTOR's being 17.6 and 14.0, respectively. When there were two causal SNPs simulated, MsCAVIAR's causal sets were smaller on average than PAINTOR's, and the difference increased when three causal SNPs were simulated. When two causal SNPs were simulated, MsCAVIAR's mean

and median set sizes were 41.6 and 36.5, respectively, while PAINTOR's mean and median set sizes were 46.4 and 45.5, respectively. Finally, with three causal SNPs, MsCAVIAR had mean and median set sizes of 52.4 and 50.0, respectively, and PAINTOR's were 60.1 and 64.5, respectively.

As the goal of most statistical fine mapping methods is to prioritize variants for functional follow-up, it lends the question of how informative a variant's posterior probability is to its causal status. We, therefore, sort the SNPs in descending order of posterior probability to determine on average how many SNPs are added to the causal set before the causal SNPs are placed in the causal set. We evaluated this quantity for MsCAVIAR, PAINTOR, and CAVIAR run on the Asian and European populations (Fig 3.2 sub-figure C). MsCAVIAR and PAINTOR were generally better at prioritizing variants than CAVIAR, again highlighting the importance of utilizing multiple studies when possible. On average, MsCAVIAR was able to capture the causal variant(s) with fewer SNPs than PAINTOR.

3.2.3 Fine mapping of high density lipoprotein across biobanks

In order to evaluate the performance of MsCAVIAR on real data, we performed a trans-ethnic, trans-biobank fine mapping analysis of HDL using summary statistics from the UK Biobank (UKB) and Biobank Japan (BBJ) projects [73, 111, 131, 147]. These studies involved 361,194 and 70,657 people, respectively. The UKB summary statistics, obtained from the Neale lab, were generated using only White Britons while Biobank Japan contained Japanese individuals [111].

To generate loci for fine mapping, we centered 1 megabase windows around genome wide-significant peak SNPs (p-value $\leq 5 * 10^{-8}$), discarding all SNPs that did not reach even marginal significance ($p > 0.05$), as they were highly unlikely to be informative and would slow down analyses. We also excluded all loci with fewer than ten SNPs in each study after filtering SNPs with $p > 0.05$, as fine mapping may not be seen as necessary or may even be trivial for existing methods when there are only a few strongly associated SNPs. Two very large loci were excluded for computational reasons. We excluded loci from chromosome six,

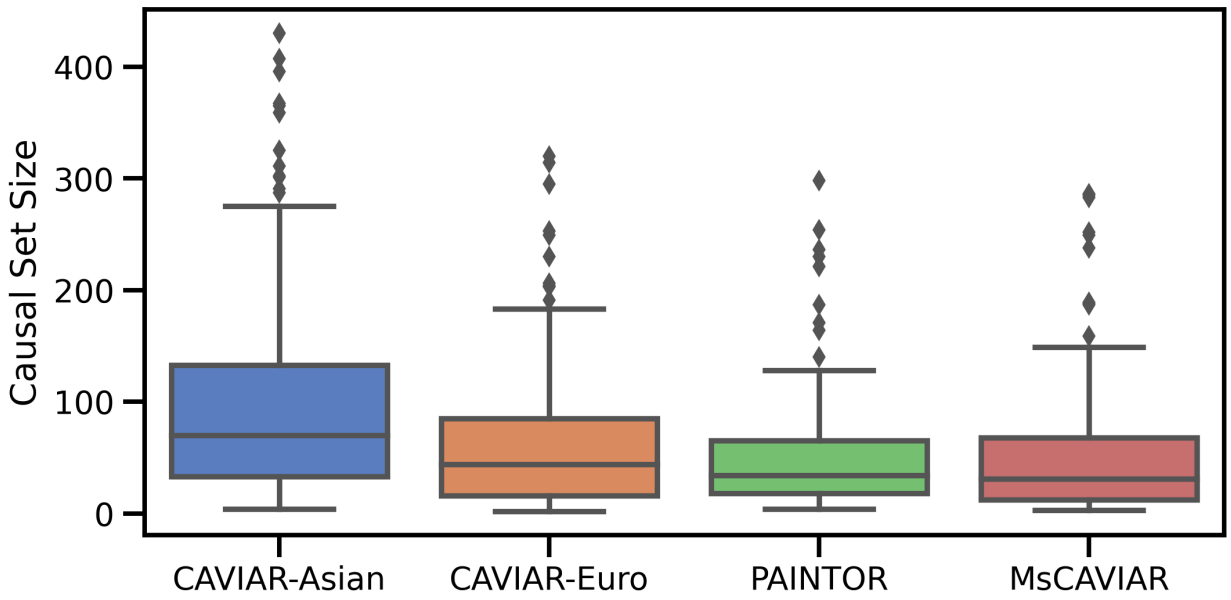


Figure 3.3: **Comparing fine mapping resolution in trans-ethnic HDL analysis.** Comparison of the results of MsCAVIAR when applied to 185 loci from two high-density lipoprotein (HDL) GWAS, White Britons from the UK Biobank and East Asian individuals from Biobank Japan, versus trans-ethnic PAINTOR and applying CAVIAR to each population individually. The y-axis is the size of the causal set for each locus. The boxes represent the interquartile range of causal set sizes identified by each tool, the lines inside the boxes represent the median, and the whiskers extend to the non-outlier extremes. Outliers are represented as dots above or below the whiskers.

where there were numerous statistically significant SNP effect sizes due to the presence of human leukocyte antigen (HLA) regions.

The procedures described above yielded 185 loci consisting of 29,479 SNPs in total. Individual locus sizes ranged from 11 to 755 SNPs. All but two SNPs in the loci had a minor allele frequency of at least 1% at least one of the studies. Linkage disequilibrium (LD) matrices were generated from the 1000 Genomes project, for both European and East Asian populations, using a preprocessing script from PAINTOR [1, 76]. We used the 1000 Genome project to generate LD to reflect the common situation where summary statistics are available but not the full genotyped data [13, 68].

We ran CAVIAR, the trans-ethnic mode of PAINTOR, and MsCAVIAR on these loci, and evaluated their causal set sizes, since these methods have been shown to be well-calibrated

and no ground truth was available (Fig 3.3). For MsCAVIAR, we set the heterogeneity parameter τ^2 to its default value of 0.52 (see Methods). For CAVIAR, we evaluated its performance when applying it to only the East Asian (BBJ) data and to only the European (UKB) data. For all methods, we set the posterior probability threshold $\rho*$ to 95% and set the maximum number of causal SNPs to 3.

While the original loci totaled 29,479 SNPs, averaging 159.3 SNPs per locus, the causal sets returned by MsCAVIAR totaled 9,390 SNPs, averaging 50.8 SNPs per locus with a median of 31 SNPs. Meanwhile, PAINTOR's causal sets totaled 9,118 SNPs (49.3 average, 34 median), CAVIAR's sets using the UKB data totaled 11,538 SNPs (62.4 average, 44 median), and CAVIAR's sets using the BBJ data totaled 18,520 SNPs (100.0 average, 70 median). Thus, similarly to our simulation study's findings, MsCAVIAR and PAINTOR generally returned smaller causal set sizes than CAVIAR, and MsCAVIAR's median causal set size was slightly smaller than PAINTOR's. In contrast with the simulation study, MsCAVIAR's average causal set size was slightly larger than that of PAINTOR's.

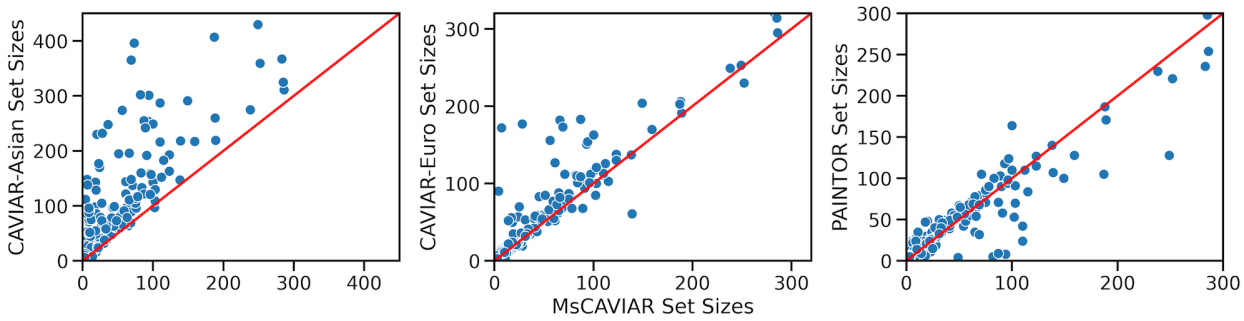


Figure 3.4: **Comparison of methods' set sizes for each locus in the trans-ethnic HDL analysis.** Comparison of the returned causal set sizes of MsCAVIAR when applied to two high-density lipoprotein (HDL) GWAS, White Britons from the UK Biobank and East Asian individuals from Biobank Japan, versus trans-ethnic PAINTOR and applying CAVIAR to each population individually. In each scatter plot, each point reflects a specific locus, and the x-coordinate is MsCAVIAR's returned causal set size, while the y-coordinate is a different method's causal set size. Diagonal lines representing equal set sizes were plotted for each scatter plot. Points above the line represent loci where the alternate method had a larger causal set size than MsCAVIAR, while points below the line indicate the opposite.

As an additional way of viewing the results, we generated scatter plots of the causal set sizes at each locus for MsCAVIAR compared to those of PAINTOR and CAVIAR (Fig 3.4).

This visualizes the comparative causal set sizes at individual loci. The scatter plots and their associated lines of equality reveal that MsCAVIAR's set sizes were consistently smaller than CAVIAR's across almost all loci, with one notable exception in which CAVIAR's causal set size was substantially smaller than MsCAVIAR's. The comparison with PAINTOR illustrates how MsCAVIAR's median causal set size was smaller than PAINTOR's but its average was higher: MsCAVIAR returned slightly smaller causal set sizes than PAINTOR for most loci, but in some cases, MsCAVIAR's causal set size was much larger than PAINTOR's, dragging MsCAVIAR's average causal set size above that of PAINTOR.

3.2.4 MsCAVIAR is well-calibrated when sample sizes differ between studies

We begin by noting that SuSiE takes a different approach to fine mapping from the other methods [158]. Instead of returning a causal set, SuSiE returns (potentially multiple) credible sets for a locus, each of which is expected to contain at least one causal SNP. The goal of SuSiE is not to capture all causal variants in a locus, but to return one or more minimal size credible sets, each of which has ρ probability of containing at least one true causal effect. This explains why SuSiE is not well-calibrated according to our causal set definition, which expects all casual variants to be captured with probability ρ^*. It is worth noting, however, that SuSiE's credible set is equivalent to the causal set (as defined by the other methods) when the methods assume that there is only one causal SNP in a locus. With this caveat with SuSiE in mind, we state that the inclusion of SuSiE was for completeness and omit the method from further discussion.

We conducted this set of simulations because studies often have different sample sizes. When this occurs, the non-centrality parameters of their SNPs will differ proportional to the sample size in addition to heterogeneity. We tested whether MsCAVIAR would still be well-calibrated in this setting, and compared it again with trans-ethnic PAINTOR as well as with CAVIAR and SuSiE run on the individual populations (Fig F). In order to evaluate performance under this scenario, we used two regions from the 1000 Genomes project to generate LD matrices for the SNPs at that locus for both European and East Asian

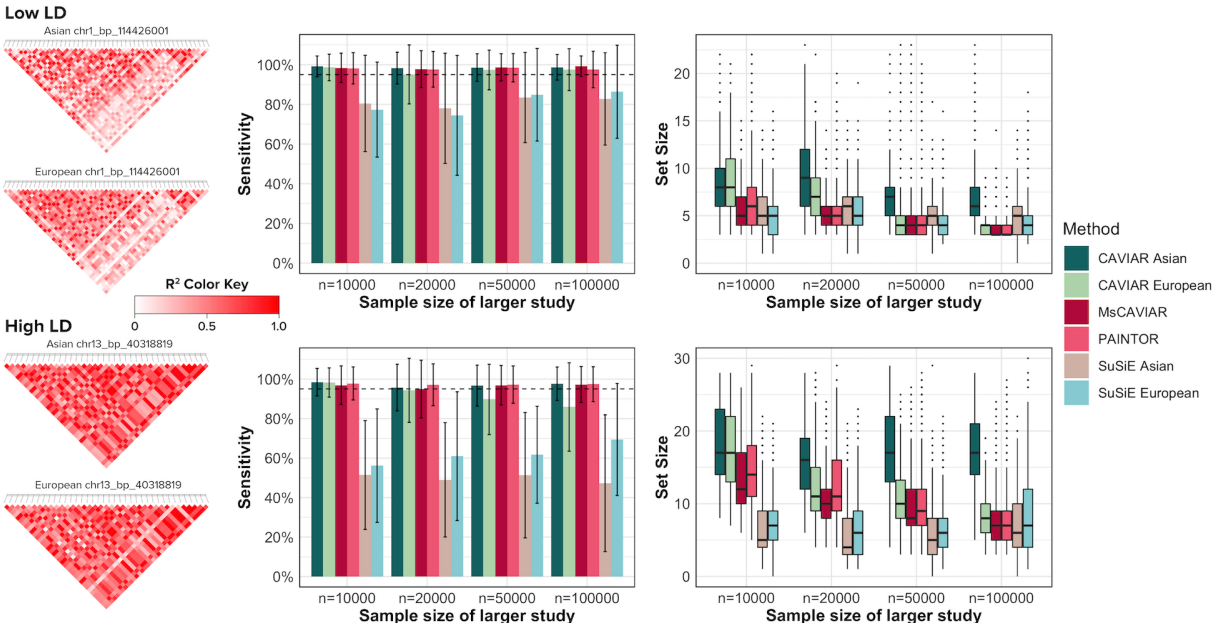


Figure F: **Comparison of sensitivity and set size using simulated studies with unequal sample sizes.** Comparison of the methods with 3 causal variants implanted and imbalanced sample sizes. The size of the Asian population was fixed at 10,000, while the European study was set to be 1, 2, 5, or 10 times larger. Both low LD (top half) and high LD (bottom half) settings were evaluated. The bar plots (left) display the sensitivity of the methods, with standard deviation bars included. The dashed line reflects the expected posterior probability of recovering all causal SNPs; methods that reach this threshold are considered well-calibrated. The box plots (right) show the set sizes returned by the methods; for SuSiE, this is calculated as the sum of the sizes of credible sets returned. The boxes represent the interquartile range of causal set sizes identified by each tool, the lines inside the boxes represent the median, and the whiskers extend to the non-outlier extremes. Outliers are represented as dots above or below the whiskers. SuSiE's credible sets differ from the causal sets of the other methods in that SuSiE does not attempt to capture all causal SNPs, so the sensitivity calibration is not directly comparable to the other methods.

populations [1]. Out of these loci, we selected one region with relatively low LD, where 20% of the SNP pairs have LD equal to or higher than 0.5, and one region with relatively high LD, where 80% of the SNP pairs have LD equal to or higher than 0.5. These represent easier and more difficult scenarios, respectively, for fine mapping, since LD makes signals more difficult to distinguish. We pruned groups of SNPs that were in perfect LD in one or more of the populations, leaving one SNP for each. After pruning, the low LD matrix contained 48 SNPs and the high LD matrix contained 38 SNPs. Using these LD matrices, we implanted

causal SNPs and simulated their non-centrality parameters. In each simulation, we implanted either 1, 2, or 3 causal SNPs. Each causal SNP's true non-centrality parameter Λ was drawn according to $\mathcal{N}(5.2, 0.125^2)$. We then drew the non-centrality parameter Λ_i for each study i according to $\Lambda_i \sim \mathcal{N}(\Lambda, 0.5)$, and subsequently the summary statistics S_i for each study i according to $S_i \sim \mathcal{N}(\Lambda_i \Sigma_i, \Sigma_i)$. For each number of causal SNPs, we performed 1000 replicate simulations (e.g. re-drawing the causal SNP non-centrality parameters and re-picking the causal SNPs). These simulations assumed a sample size of 10,000 individuals and we set the East Asian non-centrality parameter accordingly. We then scaled the European study's summary statistics assuming a population size 1, 2, 5, or 10 times larger (see Methods). For the sake of sufficient statistical power, we ensured that the causal variants in the smaller study were still statistically significant genome-wide. 1000 simulation replicates were run for each LD setting. In each simulation, we implanted three causal SNPs and simulated their effect sizes, with the association statistics of non-causal SNPs being based on their correlation with causal SNPs (see Methods). All methods were run with posterior probability threshold $\rho^* = 0.95$, so methods with 95% or higher sensitivity were considered well-calibrated (dashed line in the bar plots). MsCAVIAR was run with its heterogeneity parameter set at $\tau^2 = 0.5$ (see Methods).

Once again, MsCAVIAR was well-calibrated and generally returned the smallest causal set sizes. As the sample size differences grew, the difference between MsCAVIAR, CAVIAR on Europeans, and PAINTOR tended to diminish. This is likely due to the fact that we required SNPs to be genome-wide significant in the smaller study, such that the larger study had very large effect sizes for causal SNPs when there was a significant sample size imbalance, making the fine mapping problem easier. Reinforcing this interpretation is the fact that CAVIAR on Asians had consistently larger causal set sizes than the other methods when the sample size imbalance was large. All methods (exempting SuSiE) were well-calibrated in the low LD setting, but we observed that as the sample size increases with high LD that CAVIAR's calibration on the larger population decreases. This is likely due to the extremity of the situation, with exceptionally large effect sizes in combination with the high LD setting.

We again, note that SuSiE's miscalibration is due to fundamental differences between SuSiE and the other methods.

3.2.5 MsCAVIAR is robust to adjustments to the heterogeneity parameter

To examine the effect of having a mismatch between true heterogeneity and the model's parameter on MsCAVIAR, we first simulated our studies with different true heterogeneity τ^2. We used two regions from the 1000 Genomes project to generate a high LD matrix and low LD matrix as described above [1]. Using these two matrices LD matrices, we implanted causal SNPs and simulated their effect sizes. In each simulation, we implanted either 1, 2, or 3 causal SNPs. Each casual SNP's true overall non-centrality parameter Λ was drawn according to $\mathcal{N}(5.2, 0.125^2)$. The study specific non-centrality parameter Λ_i for each study i was drawn according to $\Lambda_i \sim \mathcal{N}(\Lambda, \tau^2)$, where $\tau^2 = 0.5$, 1, or 2. For each model configuration, we performed 1000 replicate simulations (e.g. re-drawing the causal SNP effect sizes and re-picking the causal SNPs). We then ran MsCAVIAR with different modeled heterogeneity settings, $\tau^2 = 0.5$, 1, or 2, on the simulations with the various true heterogeneity settings (Fig 3.6). MsCAVIAR was well-calibrated and maintained similar set sizes even when the modeled heterogeneity did not match the true heterogeneity, indicating that MsCAVIAR is fairly robust to small misspecifications of the τ^2 parameter.

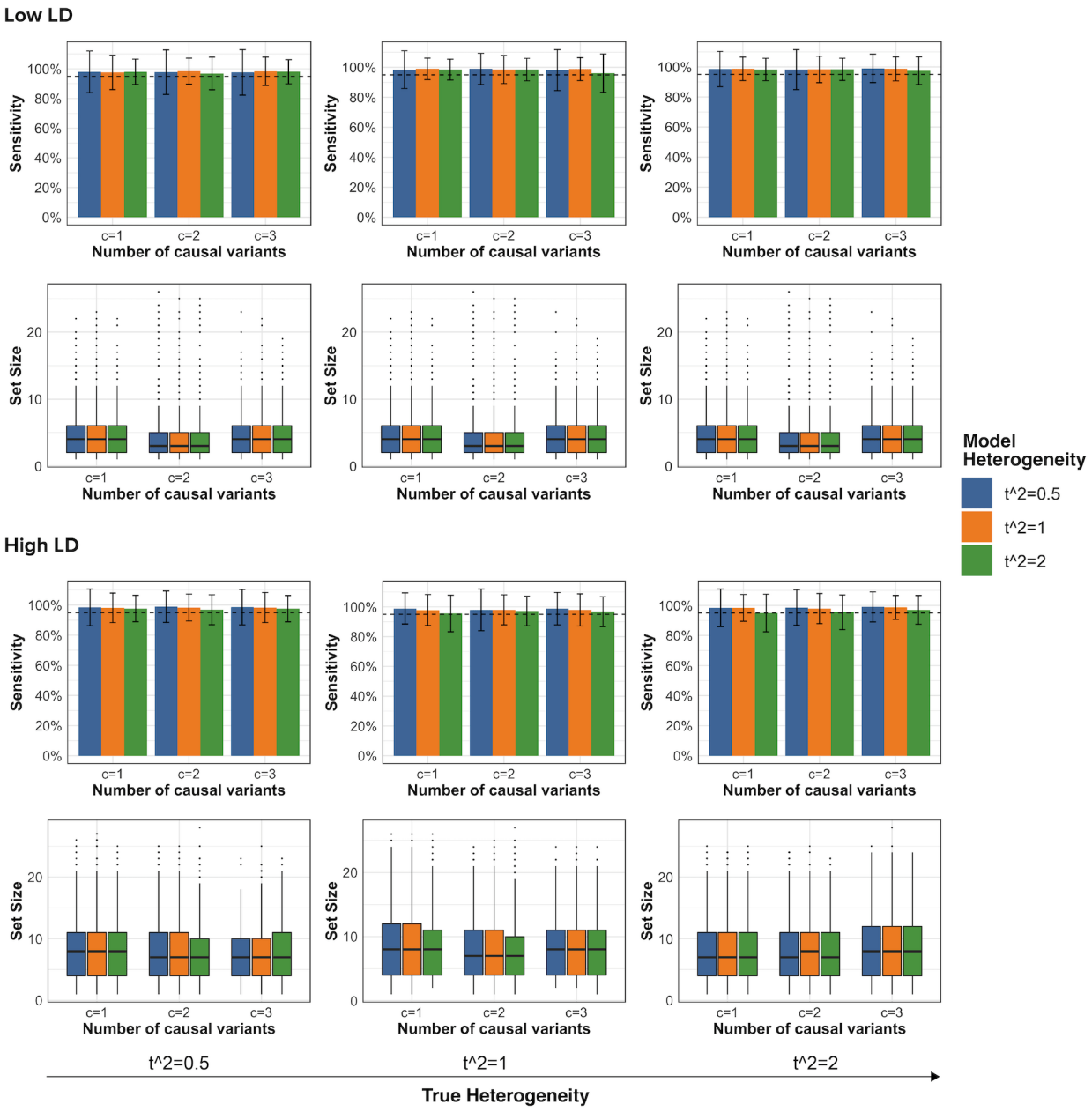


Figure 3.6: **Evaluation of the sensitivity and set sizes of MsCAVIAR results under misspecified heterogeneity parameters.** Each column of plots shows a different true heterogeneity value τ^2 used to simulate z-scores of causal variants. Different colored bars/boxes correspond to different values of τ^2 used internally in MsCAVIAR's model, referred to as the Model Heterogeneity. The model is misspecified when the Model Heterogeneity does not match the True Heterogeneity. The first two rows of plots are based on a low LD locus, and the bottom two rows are based on a high LD locus. The bar plots (1st and 3rd rows) display the sensitivity of the results, with standard deviation bars included. The dashed line reflects the expected posterior probability of recovering all causal SNPs; methods that reach this threshold are considered well-calibrated. The box plots (2nd and 4th rows) show the set sizes returned by MsCAVIAR. The boxes represent the interquartile range of causal set sizes identified by each tool, the lines inside the boxes represent the median, and the whiskers extend to the non-outlier extremes. Outliers are represented as dots above or below the whiskers. Simulations were performed with c=1, c=2, or c=3 causal variants.

3.2.6 Out-of-sample LD matrices degrade the accuracy of fine mapping

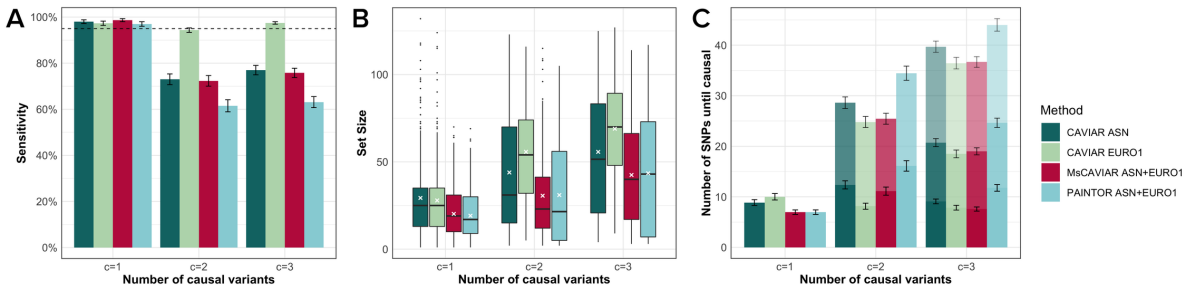


Figure 3.7: **Comparison of sensitivity, precision, and set sizes using simulated data and out-of-sample LD matrices.** We compare MsCAVIAR, PAINTOR, and CAVIAR with $c \in \{1, 2, 3\}$ causal variants averaging over 3 loci and 5 levels of heritability with 20 replicates for each value of c. (A) Bar graph indicating the sensitivity of the method and the expected posterior probability, ρ, of recovering all causal SNPs represented as a dashed line. (B) Box plots showing the average set sizes each method returns. Each box is the interquartile range of causal set sizes. The middle black line represents the median and the white crosses indicating the mean. (C) Bar graph displaying the average number of SNPs in descending order of posterior inclusion probability (PIP) until 1,2, or 3 causal SNPs are identified. Stacked bars represent an increasing number of causal SNPs identified until the true number of causal SNPs (x-axis) are identified.

The methods CAVIAR, PAINTOR, and MsCAVIAR are designed such that they only require the z-scores as summary data and may use an external LD matrix representative of the samples for fine mapping. Previous work has shown that this approach is underpowered [85]. In section 3.2.2 we provide each fine mapping method the LD matrices generated in-sample. Here, we simulate data in an almost identical fashion, but we generate the input LD matrices using the 1000 Genomes Project to demonstrate the impact of the out-of-sample LD matrices on our fine mapping results. The only difference was that we provide the LD matrices generated from the 1000 Genomes samples to the methods instead of their in-sample LD matrices.

For the results shown in Fig 3.7, we use two sets of 9,000 unrelated individuals from the UK Biobank with European and Asian ancestry and then identify the corresponding populations in 1000 Genomes Project to generate the out-of-sample LD matrices. For the 9,000 UK Biobank samples with European ancestry, we select the 503 samples in the super population "EUR" in the 1000 Genomes Project as the reference sample. For the individuals with

Asian ancestry, we needed to use two super populations "SAS" and "EAS" due to the UK Biobank sample containing 1600 individuals with Chinese ancestry, 5900 individuals with Indian ancestry, and 1800 individuals with other Asian ancestry. We generate our representative sample using all 489 "SAS" individuals and 123 "EAS" individuals sampled across sub-populations. This sub-sampling was done to approximate the proportion of individuals with Chinese and Indian ancestry. We note that while our example of out-of-sample LD for Asian ancestry is more extreme than when only 1 super population is used, it highlights how the accuracy of the out-of-sample LD impacts fine mapping.

In the results shown in Fig 3.7, we compare MsCAVIAR to trans-ethnic PAINTOR and to CAVIAR run on the Asian and European populations, separately. We average the results over the set of simulations for each number of causal SNPs: 1, 2, and 3. All methods were run using $\rho^* = 0.95$ as the posterior probability threshold; therefore methods were considered well-calibrated if their sensitivity was at least 95% (dashed line in Fig 3.7 sub-figure A), and we set MsCAVIAR's heterogeneity parameter to $\tau^2 = 0.52$ (see Methods). We evaluated the precision of the methods in Fig 3.7 sub-figure B where we show the causal set size. This metric is informative when the method returns the causal variant(s) because then fewer non-causal variants or false positives are being returned in the set.

When there is only 1 causal SNP, all methods are well-calibrated; however, only CAVIAR when analyzing European samples is well-calibrated when there are 3 causal SNPs and is slightly below the threshold for 2 implanted causal variants (94.3%). All other methods see a serious degradation in their sensitivity. This decrease in performance is a result of a poor approximation to the LD matrix for individuals with Asian ancestry. When the out-of-sample LD accurately reflects the sample, as is the case for European ancestry, CAVIAR returns results comparable to when an in-sample LD matrix is provided. For MsCAVIAR and PAINTOR, however, we see incorporating two populations does not help when one sample's LD is poorly approximated. Though the set sizes are smaller in Fig 3.7, the specificity is also lower than either run of CAVIAR. While MsCAVIAR and PAINTOR are both poorly calibrated, we see that MsCAVIAR is more robust to the out-of-sample LD than PAINTOR.

Further work would need to be done to explore this phenomenon.

At present, we encourage users to use the in-sample LD matrix whenever possible. If this is not a possibility, we advise the user to interpret their results with the understanding the out-of-sample LD may fail to provide well-calibrated results, and the quality of results depend how well the out-of-sample LD approximates the in-sample LD. Future work could also enable the method to incorporate the sufficient summary data described in [85].

3.3 Discussion

In this chapter, we introduced MsCAVIAR, a method for identifying causal variants in associated regions while leveraging information from multiple studies. Our approach requires only summary statistics as opposed to genotype data and handles heterogeneity of effect sizes, differing sample sizes, and different LD structures between studies, making trans-ethnic fine mapping an ideal application. We demonstrated that our method is well-calibrated and improves fine mapping resolution in simulation studies.

We make several important assumptions in this model, which may not always be true. It has been shown that many causal SNPs are shared across populations [83, 101, 120]. MsCAVIAR is designed to leverage this phenomenon for increased power; however, causal variants may be unique to one population. In those instances, MsCAVIAR's model doesn't match the data, so it may not be well-calibrated or it may return large causal sets. If one population has an obvious GWAS signal while the other population(s) lack even a marginally significant signal in the same locus, applying CAVIAR to the population with signal may be more appropriate.

We also assume that all studies are drawn with equal heterogeneity τ^2. This is unlikely to be true if multiple studies are from a single population while another study is from a different population. In such a scenario, we recommend grouping the studies by population, running fixed effects meta-analysis on each group, and then running MsCAVIAR on the results for the different groups. Concretely, the input summary statistics for MsCAVIAR should be

the results from the meta-analysis of each population, and the input LD matrices should be derived from either the genotype data (if available) or the appropriate reference panels for each population. However, it is still possible that even ostensibly different populations may be more similar to each other at certain loci than other populations. Therefore, we plan to extend our method to handle this case in future work.

In practice, we set the τ^2 parameter to a fixed value, which was chosen to give power to detect both small and large amounts of heterogeneity (see section 3.4.7). This value could, in principle, be adjusted based on the apparent heterogeneity present in the data. However, care would have to be taken to not overfit the parameter to the summary statistics in each locus, since the heterogeneity of different causal SNPs can vary across loci and some causal SNPs may be missed when the heterogeneity parameter is overfitted. Future work could develop a procedure for fitting this parameter.

Several methodological extensions to MsCAVIAR are possible as well. MsCAVIAR aims to return a causal set that contains all causal SNPs in a locus, while another fine mapping method, SuSiE solves a complementary problem by returning one or more credible sets that each contain at least one causal SNP [158]. The advantage of the former approach is its completeness in terms of identifying all causal signals, while the advantage of the latter approach is its ability to separate distinct causal signals within a locus into separate sets. A future extension to MsCAVIAR could aim to accomplish the benefits of both by returning a causal set with all causal SNPs, and then partitioning this set into distinct subsets with separate causal signals.

Functional information can in principle be factored into MsCAVIAR's model by modifying the prior distribution $P(C)$ so that not every variant has the same prior probability of being causal, as described in the CAVIAR paper [68]. However, setting these priors arbitrarily can yield misleading results, and future work is needed to determine how best to model various functional priors in the context of MsCAVIAR's model.

Finally, stochastic search could be used to speed up MsCAVIAR in cases where there are possibly many causal variants [7, 13]. MsCAVIAR's runtime is largely determined by the

number of SNPs in the locus and the number of causal SNPs allowed: if there are M total SNPs and up to K are allowed to be causal, then there are potentially up to $\binom{M}{K}$ causal status vectors to evaluate. Thus, runtime can become an issue when there are many SNPs in a locus or many studies, and especially when users desire to allow for more than three possibly causal SNPs at a locus. Stochastic search can help reduce the search space by not evaluating every possible combination of causal SNPs, though this involves managing the risk of missing the optimally minimal causal set.

3.4 Materials and Methods

3.4.1 Fine mapping in a single study (CAVIAR)

We now describe a standard approach for fine mapping significant variants from a genome-wide association study (GWAS). In the GWAS, let there be N individuals, all of whom have been genotyped at M variants. For each individual n, we measure a quantitative trait y_n, resulting in the $N \times 1$ column vector Y of phenotypic values. We denote G as the $N \times M$ matrix of the genotypes where $g_{nm} \in \{0, 1, 2\}$ is the minor allele count for the nth individual at variant m. We standardize G according to the population proportion p of the minor allele and denote this as X where $x_{ij} \in \{\frac{-2p}{\sqrt{2p(1-p)}}, \frac{1-2p}{\sqrt{2p(1-p)}}, \frac{2-2p}{\sqrt{2p(1-p)}}\}$.

We assume Fisher's polygenic model, which means Y is normally distributed and each variant x_m has a linear effect on Y. We, therefore, have the following model:

$$Y = \mu 1 + \sum_{m=1}^{M} \beta_m x_m + e \tag{3.1}$$

where β_m is the effect size of variant x_m and e is the variation in Y not explained by additive genetic effects and follows the Gaussian distribution $e \sim \mathcal{N}(0, \sigma_e^2 I)$.

We now model the observed summary statistics $S = [s_1, \ldots, s_m]$ according to

$$S|\Lambda_C \sim \mathcal{N}(\Sigma\Lambda_C, \Sigma) \tag{3.2}$$

where Σ represents the pairwise Pearson correlations between the genotypes. $\Lambda_C = [\lambda_{C_1} \dots \lambda_{C_M}]$ represents the true standardized causal effect sizes of each SNP, where each entry $\lambda_{C_m} = 0$ if SNP m is non-causal and $\lambda_{C_m} \neq 0$ otherwise.

The distribution of Λ_C can be defined as:

$$\Lambda_C | C \sim \mathcal{N}(0, \Sigma_C) \tag{3.3}$$

where $C = \{0,1\}^M$ is an $M \times 1$ binary vector indicating whether each variant is causal, and

$$\Sigma_C = \begin{cases} 0, & \text{if i} \neq \text{j.} \\ \sigma^2, & \text{if } i \text{ is causal.} \\ \epsilon, & \text{if } i \text{ is not causal.} \end{cases} \tag{3.4}$$

and where ϵ is a small constant to ensure that the matrix Σ_C is full rank. (We later relax the need for Σ_C to be full rank in "Handling Low Rank LD Matrices"). Here, and below, we use the shorthand σ^2 to represent the variance of the λ_{C_m} (see section 3.4.6 for details on this parameter). The off-diagonals of Σ_C are zero because the effect sizes of causal variants are independent of one another.

We use the shorthand $\Lambda = \Sigma\Lambda_C$ to refer to the non-centrality parameters (NCPs) of the statistics of all SNPs, which are induced by Linkage Disequilibrium (LD) with the causal SNPs. Thus, $S|\Lambda \sim \mathcal{N}(\Lambda, \Sigma)$. Since $\Lambda = \Sigma\Lambda_C$ and LD structure is symmetric ($\Sigma = \Sigma^T$), we have the following distribution for $\Lambda|C$:

$$(\Lambda | C) \sim \mathcal{N}(0, \Sigma\Sigma_C\Sigma) \tag{3.5}$$

We will now define γ as the probability of a variant being causal, which makes the causal status for the mth variant a Bernoulli random variable with the following probability mass function: $f(c_m; \gamma) = \gamma^{c_m}(1-\gamma)^{1-c_m}$. We assume the causal status for each variant is

independent of the other variants, leading to the following prior for the our indicator vector: $P(C) = \prod_{m=1}^{M} \gamma^{C_m}(1-\gamma)^{1-C_m}$. Assuming that each variant has a probability γ of having a causal effect, the prior can then be written as follows:

$$P(\Lambda, C) = P(\Lambda|C)P(C) = f(\Lambda, 0, \Sigma_C) \prod_{m=1}^{M} \gamma^{C_m}(1-\gamma)^{1-C_m} \tag{3.6}$$

where $f(\Lambda, 0, \Sigma_C)$ is the probability density function shown in equation 3.5.

We determine which variants are causal by calculating the posterior probability of each configuration $C^* \in \mathcal{C}$, where $\mathcal{C}$ is the set of all possible configurations, given the set of summary statistics:

$$P(C^*|S) = \frac{P(S|C^*)P(C^*)}{\sum_{c\in\mathcal{C}} P(S|c)P(c)} = \frac{\int_{\Lambda_{C^*}} P(S|\Lambda, C^*)P(\Lambda = \Sigma\Lambda_{C^*}, C^*)d\Lambda_{C^*}}{\sum_{c\in\mathcal{C}} \int_{\Lambda_c} P(S|\Lambda, c)P(\Lambda = \Sigma\Lambda_c, c)d\Lambda_c} \tag{3.7}$$

For us to calculate the posterior probability of C^* given S, we need to integrate over all possible values for the non-centrality parameters of the causal variants in Λ in order to get the values of Λ that makes observing S most probable.

3.4.2 Conjugate priors enable efficient modeling of likelihood functions

The integral above is intractable in the absence of parametric assumptions about the data. Fortunately, a closed-form solution is available due to the fact that, when a conjugate prior is multivariate normally distributed, its predictive distribution is also multivariate normal. As shown above, $S|\Lambda \sim \mathcal{N}(\Lambda, \Sigma)$ and $(\Lambda|C) \sim \mathcal{N}(0, \Sigma\Sigma_C\Sigma)$. The predictive form of S is then

$$S \sim \mathcal{N}(0, \Sigma + \Sigma\Sigma_C\Sigma) \tag{3.8}$$

However, computing the likelihood of S with this distribution is still computationally expensive. Consider the multivariate normal probability density function, assuming the

variable Z below is MVN distributed with mean μ and covariance matrix Σ:

$$f(Z;\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)^M|\Sigma|}} exp(-\frac{1}{2}(Z-\mu)^T\Sigma^{-1}(Z-\mu)) \tag{3.9}$$

For S, the covariance matrix is $\Sigma+\Sigma\Sigma_C\Sigma$, which has dimension $(M \times M)$, where M is the number of SNPs in each study. Taking the determinant or inverse of this covariance matrix, as required by the above likelihood function, would take $O(M^3)$ time. Here, we demonstrate how to compute this likelihood efficiently, leveraging insights from several studies that have explored this topic [13, 25, 94].

We need to compute $S^T(\Sigma + \Sigma\Sigma_C\Sigma)^{-1}S$ and $|\Sigma + \Sigma\Sigma_C\Sigma|$ (note that our μ is 0). We can factor out Σ from both of the equations above:

$$S^T(\Sigma + \Sigma\Sigma_C\Sigma)^{-1}S = S^T\Sigma^{-1}(I + \Sigma_C\Sigma)^{-1}S \tag{3.10}$$

$$|\Sigma + \Sigma\Sigma_C\Sigma| = |\Sigma||I + \Sigma_C\Sigma| \tag{3.11}$$

Notably, $S^T\Sigma^{-1}$ and $|\Sigma|$ can be computed once and re-used for every causal configuration Σ_C. Below, we assume Σ is of full-rank; Lozano et. al show how to address the low-rank case [94].

We use the Woodbury matrix identity to speed up the matrix inversion equation [65]:

$$(A + UEV)^{-1} = A^{-1} - A^{-1}U(E^{-1} + VA^{-1}U)^{-1}VA^{-1} \tag{3.12}$$

Here, we set $A = I_{M\times M}$, $E = I_{K\times K}$ where K is the number of causal SNPs per study, and $UV = \Sigma_C\Sigma$. In particular, U is the $(M \times K)$ matrix of rows corresponding to causal SNPs in Σ_C. We are taking advantage of the fact that rows corresponding to non-causal SNPs are zeros and thus do not affect the matrix multiplication. Similarly, V is the corresponding columns of Σ, and is $(K \times M)$. Applying the Woodbury matrix identity to our case, we get:

$$
\begin{aligned}
(I_{M\times M} + \Sigma_C \Sigma)^{-1} &= (I_{M\times M} + UV)^{-1} \\
&= I_{M\times M}^{-1} - I_{M\times M}^{-1} U (I_{K\times K}^{-1} + V I_{K\times K}^{-1} U)^{-1} V I_{M\times M} \\
&= I_{M\times M} - U(I_{K\times K} + VU)^{-1} V
\end{aligned} \tag{3.13}
$$

Crucially, we are now inverting a $(K \times K)$ matrix instead of an $(M \times M)$ matrix, where $K << M$ since most SNPs are not causal [94]. We use Sylvester's determinant identity to speed up the determinant computation as follows [2]:

$$
|I_{M\times M} + UV| = |I_{K\times K} + VU| \tag{3.14}
$$

Similarly, we are computing the determinant of a $(K \times K)$ matrix instead of an $(M \times M)$ matrix. Using these speedups, the computation of the likelihood function of S is reduced from $O(M^3)$ to $O(K^3)$ plus some $O(MK^2)$ matrix multiplication operations, which is tractable under the reasonable assumption that each locus has at most $K = 3$ causal SNPs. In section 3.4.4, we discuss the computational complexity and the use of these efficient matrix computations in the multiple study setting.

3.4.3 Fine mapping across multiple studies (MsCAVIAR)

As GWAS continue to grow in size, frequency, and diversity, there is an increasing need for fine mapping methods that leverage results from multiple studies of the same trait. A simple approach is to assume that there is one true non-centrality parameter for every variant; therefore Λ_C is identical across studies. This approach is referred to as a fixed effects model. In this case, the qth study's $\Lambda_{C_q} = \Lambda_C$.

While there is evidence that many causal SNPs are shared across populations, the assumption that the true causal non-centrality vector Λ_C is the same across studies is unrealistic, especially when the studies are measured in different ethnic groups [83, 83, 101, 101, 120, 163].

We relax this assumption by utilizing a random effects model, in which each study q is allowed to have a different Λ_{Cq}. Under this model, a causal SNP m has an overall mean non-centrality parameter, which we denote with the scalar λ_{C_m}, from which the non-centrality parameter for SNP m in each study q, denoted by the scalar $\lambda_{C_{mq}}$, is drawn with heterogeneity (variance) τ^2. According to the polygenic model, λ_{C_m} is distributed as $\lambda_{C_m} \sim \mathcal{N}(0, \sigma^2)$; therefore, $\lambda_{C_{mq}}$ is distributed as $\lambda_{C_{mq}} \sim \mathcal{N}(\lambda_{C_m}, \tau^2)$. Consequently, the vector Λ_{Cm} for this SNP across all studies will have the following distribution:

$$\Lambda_{Cm} \sim \mathcal{N}(0, \sigma^2 11^T + \tau^2 I) \tag{3.15}$$

where Q is the number of studies, 1 is a $(Q \times Q)$ matrix of 1s, and I is the $(Q \times Q)$ identity matrix. Intuitively, since the SNP m was drawn with variance σ^2, this variance component is shared across studies, while the variance component τ^2 is study-specific and therefore it is only present along the diagonal of the covariance matrix. If a variant is not causal, its true effect size should be zero. We construct a matrix Λ_C of size $(MQ \times MQ)$, where M is the number of SNPs and each row corresponds to the Q-length vector Λ_{Cm} corresponding to SNP m. In practice, we ensure that this matrix is full-rank by drawing the non-causal SNPs according to $\Lambda_{Cm} \sim \mathcal{N}(0, \epsilon I)$, where ϵ is a small constant.

From this we will now build out the posterior probability of $P(C^*|S_q)$ similarly to equation 3.7. Now instead of $\Lambda_{C_q} = \Sigma_q \Lambda_C$ for study q, we have to account for $\Lambda_q = \Sigma_q \Lambda_{C_q}$ where Λ_{C_q} is drawn from a multivariate normal distribution. This means we have to integrate over the domain-space of Λ_{C_q} to as well as Λ_C to describe $P(C^*|S_q) = \frac{P(S_q|C^*)P(C^*)}{\sum_{C \in C} P(S_q|C)P(C)}$

$$P(C^*|S_q) = \frac{\int_{\Lambda_{C_q^*}} P(S_q|\Lambda_q, C^*) \int_{\Lambda_{C^*}} P(\Lambda_q = \Sigma_q \Lambda_{C_q^*}|\Lambda_{C^*}, C^*) P(\Lambda_{C^*}, C^*) d\Lambda_{C^*} d\Lambda_{C_q^*}}{\sum_{c \in C} P(S_q|\Lambda_q, c) \int_{\Lambda_{c_q}} P(S_q|\Lambda_q, c) \int_{\Lambda_c} P(\Lambda_q = \Sigma_q \Lambda_{c_q}|\Lambda_c, c) P(\Lambda_c, c) d\Lambda_c d\Lambda_{c_q}} \tag{3.16}$$

3.4.4 Cojugate priors enable efficient meta-analysis of likelihood functions

Now that we have described the distribution of each SNP in our meta-analysis, we show how to jointly analyze them. We begin by explicitly defining the structure of the covariance matrix between studies by way of a small example with three SNPs at a locus in two different studies. Since the covariance of a matrix is undefined, we denote $vec(\Lambda_C)$ as the vectorized form of the original matrix (Λ_C). Concretely:

$$\text{vec}(\Lambda_C) = \text{vec}\left(\begin{bmatrix} \lambda_{C_{11}} & \lambda_{C_{21}} \\ \lambda_{C_{12}} & \lambda_{C_{22}} \\ \lambda_{C_{13}} & \lambda_{C_{23}} \end{bmatrix}\right) = \begin{bmatrix} \lambda_{C_{11}} \\ \lambda_{C_{12}} \\ \lambda_{C_{13}} \\ \lambda_{C_{21}} \\ \lambda_{C_{22}} \\ \lambda_{C_{23}} \end{bmatrix} \tag{3.17}$$

Assume SNPs 1 and 3 are causal and SNP 2 is not causal. Then the vectorized form of the non-centrality parameters given the causal statuses has the following multivariate normal distribution:

$$(\text{vec}(\Lambda_C)|\text{vec}(C)) \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \left[\begin{array}{ccc|ccc} \sigma^2+\tau^2 & 0 & 0 & \sigma^2 & 0 & 0 \\ 0 & \epsilon & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2+\tau^2 & 0 & 0 & \sigma^2 \\ \hline \sigma^2 & 0 & 0 & \sigma^2+\tau^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & \epsilon & 0 \\ 0 & 0 & \sigma^2 & 0 & 0 & \sigma^2+\tau^2 \end{array}\right]\right) \tag{3.18}$$

We call the covariance matrix above Σ_C. Viewing Σ_C as having a block structure, the blocks along the diagonal represent SNPs from the same study, while off-diagonal blocks represent SNPs from different studies. Here Σ_C is $(3 * 2 \times 3 * 2) = (6 \times 6)$; in general, for M SNPs and Q studies, Σ_C will be $(MQ \times MQ)$. In other words, there will be an $(Q \times Q)$ grid of $(M \times M)$ blocks. Within each block, the diagonal represents each SNP's variance, while the off-diagonal represents covariation between different SNPs. As SNPs are assumed to be independent, these are always 0. There are two variance components: the global genetic variance σ^2 from which the global mean non-centrality parameter for a SNP is drawn, and the heterogeneity between studies τ^2. When a SNP is causal, its variance (its covariance with itself in the same study) will contain both variance components $(\tau^2 + \sigma^2)$, while its covariance with the same SNP in a different study will be σ^2, because they were drawn from the same overall non-centrality parameter with variance σ^2 but were drawn separately with variance τ^2.

The Σ_C above, leaving aside ϵ for now, can alternately be written in the more-compact form

$$\Sigma_C = \begin{bmatrix} \tau^2 + \sigma^2 & \sigma^2 \\ \sigma^2 & \tau^2 + \sigma^2 \end{bmatrix} \otimes \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{3.19}$$

where $\otimes$ represents the Kronecker product operator. This can be further condensed and generalized into:

$$\Sigma_C = (\tau^2 I_Q + \sigma^2 1_Q 1_Q^T) \otimes \text{diag}(1_{causal})_M \tag{3.20}$$

where Q is the number of studies, M is the number of SNPs, $1_Q 1_Q^T$ is the $(Q \times Q)$ matrix of all 1s, I_Q is the $(Q \times Q)$ identity matrix, and $diag(1_{causal})_M$ is an $(M \times M)$ diagonal matrix whose diagonal entries are given by the $(1 \times M)$ indicator vector 1_{causal} whose entries m are

1 if SNP m is causal and 0 otherwise.

As with CAVIAR, the ϵ entries along the diagonal are small numbers to ensure full rank. Also note that the CAVIAR model is a specific case of this model, in which there is only one study and thus there is no τ^2 component. The CAVIAR Σ_C has the same structure as the upper left block in the Σ_C above, when there are 3 SNPs and τ^2 is set to 0.

The efficient computation properties for the single-study case also apply to the multiple-study case. In the latter setting, the matrices that need to be inverted are $(MQ \times MQ)$ instead of $(M \times M)$, where M and Q are the number of SNPs in a locus and the number of studies, respectively. Consequently, in the Woodbury matrix identity equations, U and V are $(MQ \times KQ)$ and $(KQ \times MQ)$, respectively, where $K << M$ is the number of causal SNPs, and the matrix given by the Woodbury identity is $(KQ \times KQ)$. Sylvester's determinant identity gives a matrix of this size as well. The computation time is thus reduced from $O(M^3Q^3)$ to $O(K^3Q^3)$.

3.4.5 MsCAVIAR effectively handles low rank LD matrices

The methods described above assume that the Linkage Disequilibrium (LD) matrix is full rank, in order to invert this matrix in the process of computing the Multivariate Normal (MVN) likelihood function. In practice, this is often not the case, because SNPs are sometimes in perfect LD. This can even happen when SNPs are not in perfect LD due to many highly correlated SNPs being a linear function of each other. CAVIAR employs a method to add a small amount of random noise to the diagonal of the LD matrix to avoid this, but we found this adjustment to be insufficient to avoid the latter situation when LD matrices were sufficiently large, especially with blocks of high-LD [68].

Lozano et. al. developed a method for computing the MVN likelihood function when the LD matrix is low rank [94] . MsCAVIAR implements this method and thereby avoids the aforementioned low rank issue. We briefly describe the intuition behind the method.

Since the LD matrix Σ is positive semi-definite, it can be eigendecomposed as follows:

$$\Sigma = W\Omega W^T \tag{3.21}$$

where W is the matrix of eigenvectors, such that the i-th column of W is the i-th eigenvector of Σ, and Ω is a diagonal matrix that consists of eigenvalues of Σ where the i-th diagonal element of Σ is the i-th eigenvalue of Σ. Lozano et al. then introduce a new set of summary statistics $S' = \Omega^{-1/2}W^T S$ which, using some algebra, is shown to have the joint distribution

$$S' = \Omega^{-1/2}W^T S \sim \mathcal{N}(0, I + mB\Sigma_C B^T) \tag{3.22}$$

where I is the identity matrix, m is the number of SNPs, and $B = \Omega^{-1/2}W^T$. Since $I + mB\Sigma_C B^T$ is full rank, we can compute the likelihood function for S', even when S is not full rank.

In order to evaluate the likelihood function for our original summary statistics S, we first transform the original summary statistics S to S' via $S' = \Omega^{-1/2}W^T S$, and then apply the above procedure to evaluate the likelihood function for S'. This obviates the need for the ϵ parameter previously used to ensure full rank in the definition of Σ_C, so we now define Σ_C in the single study setting as

$$\Sigma_C = \begin{cases} 0, & \text{if i} \neq \text{j or SNP } i \text{ is not causal.} \\ \sigma^2, & \text{if SNP } i \text{ is causal.} \end{cases} \tag{3.23}$$

3.4.6 Extending MsCAVIAR to model differing sample sizes

In section 3.4.3, we discussed the MsCAVIAR model, in which the non-centrality parameters $\lambda_{C_{mq}}$ for SNP m in each study q are drawn around a global mean non-centrality parameter $\lambda_{C_m} \sim \mathcal{N}(0, \sigma^2)$ with variance τ^2, such that $\lambda_{C_{mq}} \sim \mathcal{N}(\lambda_{C_m}, \tau^2)$. We note that λ_{C_m} is itself a function of the non-standardized effect size β_m, where $\lambda_{C_m} = \frac{\beta_m \sqrt{N}}{\sigma_e}$ and $\beta_m \sim \mathcal{N}(0, \sigma_g^2)$. Thus, λ_{C_m} and its variance σ are functions of the sample size N. Since the sample size may

not be consistent across the studies, this λ_{C_m} is an oversimplification that cannot be used when different studies have different sample sizes. Below, we show how to model the $\lambda_{C_{mq}}$ for each study while taking into account possibly different sample sizes.

We will again draw the qth study's non-centrality parameter for variant m according to this model. Each study q has its own sample size N_q and environmental component σ_{e_q}, and we draw it with heterogeneity parameter τ^2 as previously defined, so

$$\lambda_{C_{mq}} \sim \mathcal{N}(\frac{\beta_m}{\sigma_{e_q}}\sqrt{N_q}, \tau^2) \tag{3.24}$$

We will now operate under the standard assumption that the trait has unit variance and variance explained by any particular SNP is small, thus $\sigma_e \approx 1$.

$$\Sigma = W\Omega W^T \tag{3.25}$$

Using our previous definition for a single study, we now have

$$\Lambda|C \sim \mathcal{N}(0, \Sigma_C) \tag{3.26}$$

where

$$\Sigma_C = \begin{cases} 0, & \text{if i} \neq \text{j or SNP } i \text{ is not causal.} \\ \sigma^2, & \text{if SNP } i \text{ is causal.} \end{cases} \tag{3.27}$$

We now define σ^2 more formally to be $\sigma_g^2 N_q$ for the qth study, in the single study setting. In the multiple study setting, when we consider our matrix

$$\Sigma_C = \begin{bmatrix} \tau^2 + \sigma^2 & \sigma^2 \\ \sigma^2 & \tau^2 + \sigma^2 \end{bmatrix} \otimes \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{3.28}$$

the σ^2 along the diagonal is defined identically to the precise single study definition; however, when modeling multiple studies, this adjustment changes the covariance between causal variant for two studies. We now define $\sigma^2 = \sqrt{N_{q1}}\sqrt{N_{q2}}\sigma_g^2$ for two studies $q1$ and $q2$ with population sizes N_{q1} and N_{q2}. Note that if two studies have the same population size N, we get the original definition of $\sigma^2 = \sqrt{N}\sqrt{N}\sigma_g^2 = N\sigma_g^2$.

3.4.7 Effective parameter setting in practice

Traditionally, the effect size $\beta \sim \mathcal{N}(0, \sigma_g^2)$ would be derived as a notion of the per-snp heritability. Here we do not define σ_g^2 as such, but rather treat it as an abstraction: we avoid making any assumptions on how heritable the given trait is and how that heritability is partitioned between loci. The way we set this parameter in practice is as a parameter for statistical power. If study $q1$ has the smallest sample size, we set this value such that $\sigma = \sigma_g^2 N_{q1} = 5.2$ for all variants. This value corresponds to the traditional genome-wide significant z-score of 5.2, for which the two-sided Wald test p-value is 5×10^{-8}, which is considered significant by (conservatively) correcting for multiple testing [117]. Then the NCP for variant m in the corresponding study $q1$ is $\lambda_{C_{q1,m}} \sim \mathcal{N}(5.2, \tau^2)$. For another study $q2$ with larger sample size, its NCP is drawn as $\lambda_{C_{q2,m}} \sim \mathcal{N}(5.2\sqrt{\frac{N_{q2}}{N_{q1}}}, \tau^2)$.

This value of σ_g^2 may not represent the actual heritability partitioning, but we set the parameter this way in our method for the practical purpose of giving MsCAVIAR power to fine map borderline significant variants in the smallest study. Similarly, we set $\tau^2 = 0.52$ by default, e.g. 10% of the value of $\sigma = \sigma_g^2 N_{q1}$, with the value chosen to give power to detect both small and large amounts of heterogeneity. We empirically observed that small

misspecifications in the heterogeneity parameter do not have a substantial adverse effect (see Fig 3.6).

CHAPTER 4

Comprehensive analysis of pan-cancer determinants of somatic mutational burden with implications for survival

4.1 Introduction

Cancer is a disease caused by germline polymorphisms, somatic alterations, and the interaction of the two, and the genetic architecture of cancer has been extensively studied through both lenses [24, 40, 47, 58, 79, 122, 123]. Previous work has elucidated familial cancer risk genes through linkage studies (e.g. BRCA1, BRCA2) as well as the polygenic, common variant contribution to cancer susceptibility through GWAS [12, 23, 38, 54, 61, 78, 82, 103, 162]. Separately, the role of somatic cancer drivers has been explored through large-scale tumor sequencing efforts [8, 57, 89, 102, 128, 161]. Through these efforts, individual somatic single nucleotide variants (SNVs) as well as the accumulation of somatic SNVs, tumor mutational burden (TMB), have been shown to be highly variable between cancer types [3, 21, 138, 170]. Additionally, increased TMB has also been linked to exogenous factors such as tobacco smoke as well as endogenous defects in DNA mismatch repair and DNA replication [40, 44, 77, 143, 159]. Recently, TMB has been shown to have a parabolic relationship with overall survival (OS) with patients at the extrema fairing better than patients with intermediate-range TMB [130]. It has also been identified as a biomarker for immune checkpoint inhibitor (ICI) response, likely through the generation of neoantigens [19, 21, 55, 132, 141, 150]. Somatic copy number variants (CNVs) have been separately linked with cancer outcomes, with focal somatic CNVs linked to proliferation while arm-

level and chromosome-level aneuploidy were correlated with immune evasion [34]. While substantial work has been done to understand the genetic underpinning of cancer, limited work has explored the interplay between germline variants and somatic burden. Here, we explore the impact of clinical features and polygenic germline features on TMB and somatic copy number burden (CNB), as well as their downstream influence on patient survival.

Both patient demographics and germline/host variation has been previously shown to correlate with somatic burden and patient outcomes [29, 30, 31, 35, 41, 64, 86, 87, 148]. Age and sex have been linked to the somatic landscape of tumors, this includes both mutational signature and somatic burden. These findings suggest that patient level biological factors shape tumor evolution such as through declining DNA damage repair which is associated with increased age [67]. Others have explored the impact of the tumor biopsy site (primary versus metastatic) on TMB and CNB measurements and found a higher mutational load in metastatic sites [9, 66, 135, 145, 173]. In addition to these clinical features, researchers have identified polygenic risk scores (PRS) associated with a number of cancer related phenotypes [90, 124, 148]. This includes somatic mutational signature which may indicate a germline influence on hormone regulation and immune response within cancer [90]. PRS have also been associated with cancer subtypes, and this association may reflect how dependent the cancer subtype is on an underlying germline background [124]. Lastly, there have been associations between PRS and TMB, which along with the paper's other findings, presents evidence of a polygenic architecture of TMB [148].

While previous studies have shown that clinical and germline features influence the somatic profile of tumors, the scope of these discoveries was limited. A particular challenge is the fact that the largest public cohort of germline and somatic data, The Cancer Genome Atlas (TCGA), exhibits systematic technical/batch effects for both the germline and somatic assays, which can lead to spurious associations [16, 28, 81, 127]. Moreover, TCGA samples were largely collected prior to the "immunotherapy era" and thus cannot be linked directly to immunotherapy outcomes. While clinical sequencing of tumors has become common and led to large pan-cancer cohorts, genome-wide germline genotyping is rarely collected for the

same patients [4, 26]. Ultimately, TCGA, now over a decade old, remains one of the few cancer cohorts mined for germline-somatic associations to date.

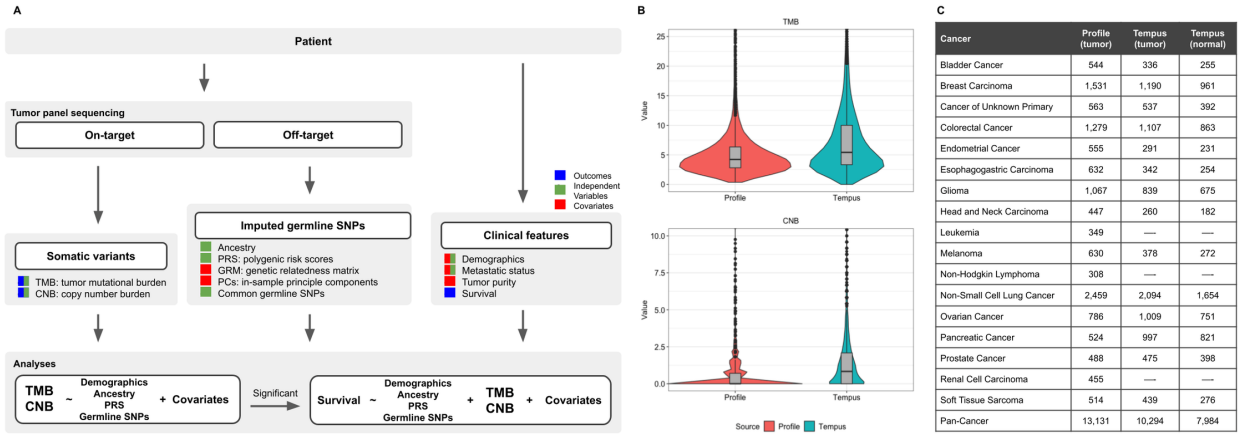
In this study, we generated a large germline-somatic cohort with ¿23,000 patients of European ancestry spanning 17 common cancers, including 1,415 treated with immunotherapy (IO). We then explored the impact of clinical features and polygenic germline features on TMB and CNB. We identified dozens significant associations, many of which were novel findings. Using 11,973 patients with treatment and survival data available, we showed that TMB/CNB as well as clinical features were associated with OS. In addition, we implicated four fine-scale genetic ancestry associations with OS in immunotherapy naive (non-IO) patients as well as one ancestry-TMB interaction effect. Separately in IO recipients, interaction analyses identified modifiers of PRS effect on CNB and TMB. Overall, we established that clinical features, fine-scale genetic ancestry and PRS shape the somatic landscape both pan-cancer and within specific cancers with implications for survival.

4.2 Results

4.2.1 Data overview

We investigated the impact of clinical and polygenic germline features on the burden of somatic alterations in two independent pan-cancer cohorts leveraging targeted tumor sequencing of $> 23,000$ patients. The two cohorts come from different institutional settings which utilized different sequencing platforms, somatic variant calling pipelines, and had varying availability of clinical features; both cohorts, however, were processed using the same bioinformatics pipeline for off-target imputation and analysis (Fig 4.1 sub-figure A).

The primary cohort, Profile, consisted of tumors sequenced during the course of routine care at Dana-Farber Cancer Institute and had extensive availability of clinical features. Each tumor was sequenced on one of three versions of the OncoPanel platform which targeted 275, 300, and 447 genes [52]. We restricted the sample to those with European ancestry (see Fig 4.2 sub-figures A and C). From this subset of samples, we removed individuals with



Cancer	Profile (tumor)	Tempus (tumor)	Tempus (normal)
Bladder Cancer	544	336	255
Breast Carcinoma	1,531	1,190	961
Cancer of Unknown Primary	563	537	392
Colorectal Cancer	1,279	1,107	863
Endometrial Cancer	555	291	231
Esophagogastric Carcinoma	632	342	254
Glioma	1,067	839	675
Head and Neck Carcinoma	447	260	182
Leukemia	349	—	—
Melanoma	630	378	272
Non-Hodgkin Lymphoma	308	—	—
Non-Small Cell Lung Cancer	2,459	2,094	1,654
Ovarian Cancer	786	1,009	751
Pancreatic Cancer	524	997	821
Prostate Cancer	488	475	398
Renal Cell Carcinoma	455	—	—
Soft Tissue Sarcoma	514	439	276
Pan-Cancer	13,131	10,294	7,984

Figure 4.1: **Overview of pipeline and cohorts**. (A) Flowchart outlining the bioinformatics pipeline for off-target imputation and analysis. Each germline, clinical, and somatic feature is color coded according to their purpose: outcomes (blue), independent variables (green), and covariates (red). (B) Distribution of somatic burden pan-cancer for both cohorts, Profile (red) and Tempus (blue). Tumor mutational burden (TMB) is shown on top and copy number burden (CNB) is depicted in the bottom panel. (C) Final sample sizes for each cancer as well as pan-cancer in Profile and Tempus with a separate column for Tempus normal samples.

microsatellite instability and selected the 17 largest cancers with 300 or more patients for a total of 13,131 tumors (Fig 4.1 sub-figure C).

The other cohort explored, Tempus, was generated in a commercial setting and contained tumors originating from multiple institutions. Tumors were sequenced on the Tempus xT next generation sequencing platform on one of three panel versions which targeted 595, 596, and 648 genes, respectively [10]. As part of its genomic profiling, Tempus collected normal-matched samples to improve the accuracy of somatic calling as shown in Fig 4.3 [10, 72]. Due to the impact of normal-matching samples on somatic calling and our non-standard imputation process, we separately evaluated the associations in Tempus using tumor and normal samples and reported any inconsistencies. Using the same protocol for inclusion (European ancestry and microsatellite stability; see Fig 4.2 sub-figures B and D), we selected 14 cancers with 200 or more individuals in a de-identified genomic database from Tempus, resulting in 10,294 tumors, 78% of which have a normal-matching sample (Fig 4.1 sub-figure C).

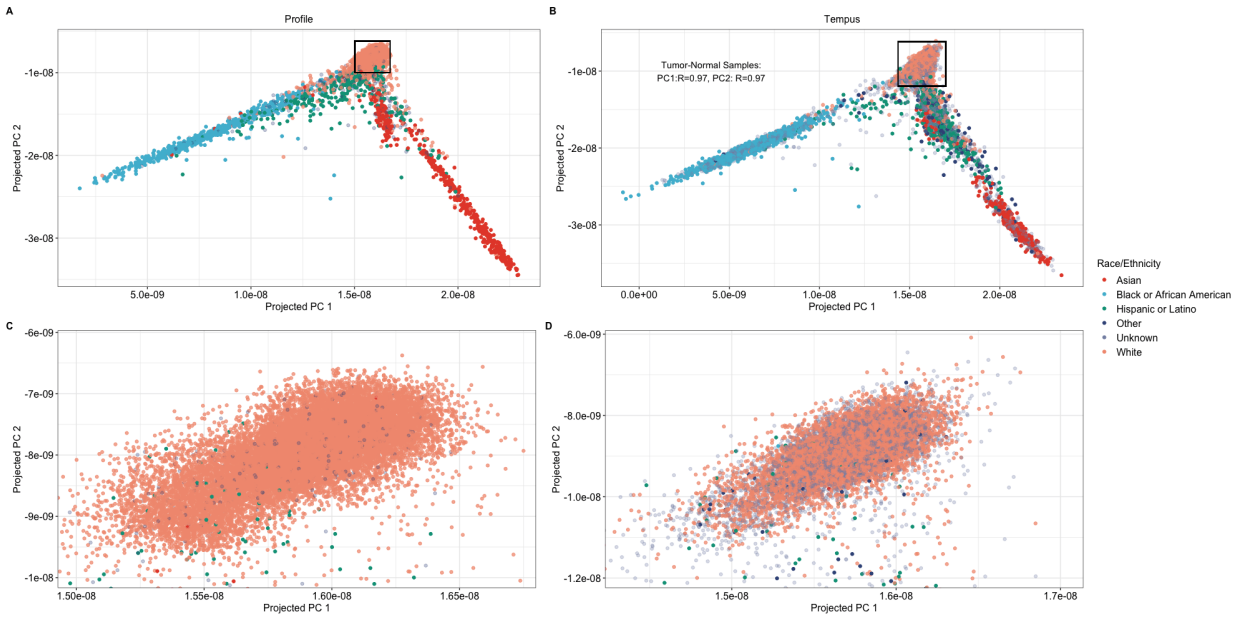


Figure 4.2: **Continental ancestry and European subset.** (A) Inferred continental ancestry in Profile, color coded by self-reported race. We restrict the analyses to individuals within two standard deviations of the mean inferred ancestry of self-reported white individuals with the boundaries shown by the black rectangle. (B) Inferred continental ancestry of Tempus tumor samples, color coded by self-reported race. A black rectangle shows the bounds of our cohort which was restricted to individuals within two standard deviations of the mean inferred ancestry of self-reported white individuals. The correlation coefficient between the tumor samples and the normal samples inferred ancestry is indicated in the top-right corner. (C) Zoomed in plot of the black rectangle in sub-figure A of Profile. (D) Zoomed in plot of Tempus black rectangle which captures European ancestry in sub-figure B.

We defined two measures of somatic burden (Fig 4.1 sub-figure B): (1) the total number of somatic single nucleotide variants (SNVs), which we call "TMB" for tumor mutational burden; (2) the total number of deep somatic copy number gains or losses, which we call "CNB" for copy number burden. In the Profile cohort, where more detailed copy number calling was available, we additionally explored a definition of CNB based on all gains or losses (rather than just deep events) which we call "All CNB". All somatic burdens were quantile normalized to adjust for their highly skewed distributions. Lastly, we created a binary variable "TMB-H" in addition to the continuous TMB phenotype, indicating whether a patient has TMB $\geq$ 10; this feature was based on the biomarker threshold for immune checkpoint inhibitor therapy [99, 100].

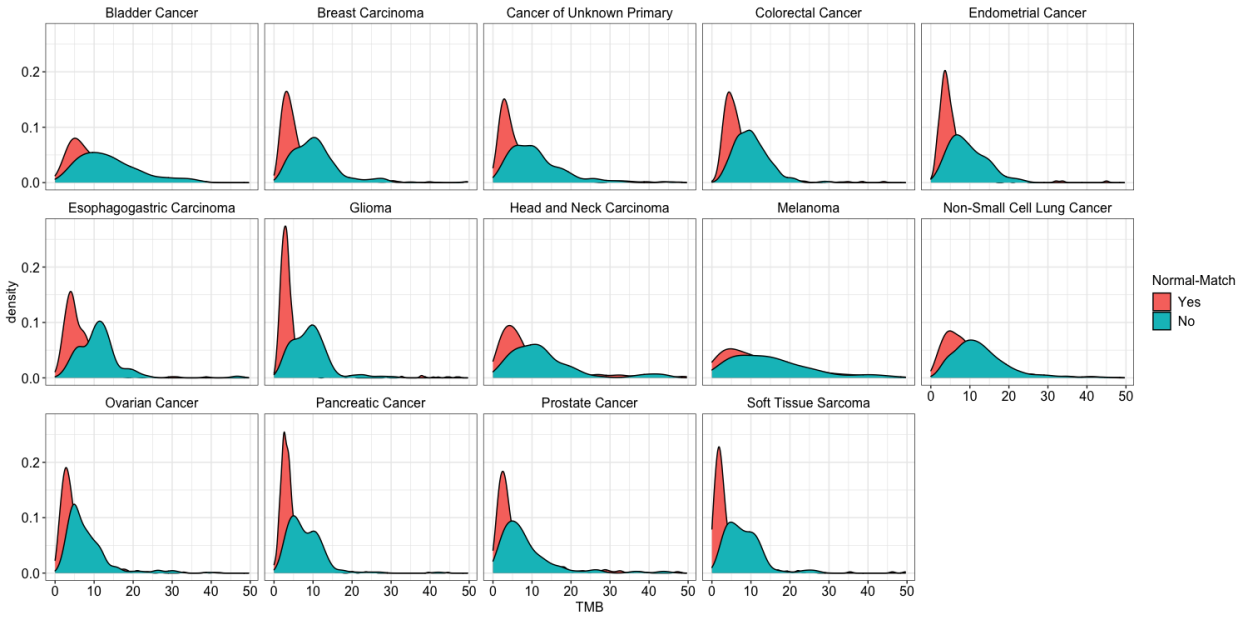


Figure 4.3: **Impact of normal-match samples on estimating TMB.** For each cancer in Tempus, we compared the distribution of TMB for individuals with a normal-matching sample (red) to those with only a tumor sample sequenced (blue).

A by-product of targeted panel sequencing is an abundance of off-target reads, which we utilized alongside the on-target reads to impute common germline variants from the 1000 Genomes Project haplotype reference panel (see Methods) [1]. With the imputed germline variants and somatic outcomes called for each individual, we were also able to analyze the polygenic impact of germline variants on TMB and CNB through PRS and fine-scale genetic ancestry.

4.2.2 Somatic burden is associated with clinical features

We first explored the association of TMB and CNB with a joint model of the broad demographic features age, sex, and metastatic status in cancer-specific and pan-cancer analyses to both provide a positive control and to explore these associations in a large, modern cohort. In total, we identified 42 significant associations in the Profile cohort at $p < 0.05$ after Bonferroni correction for multiple tests within each feature, 11 of which have been previously reported in TCGA (Table 4.1). We were able to test 26 of the 42 significant associations

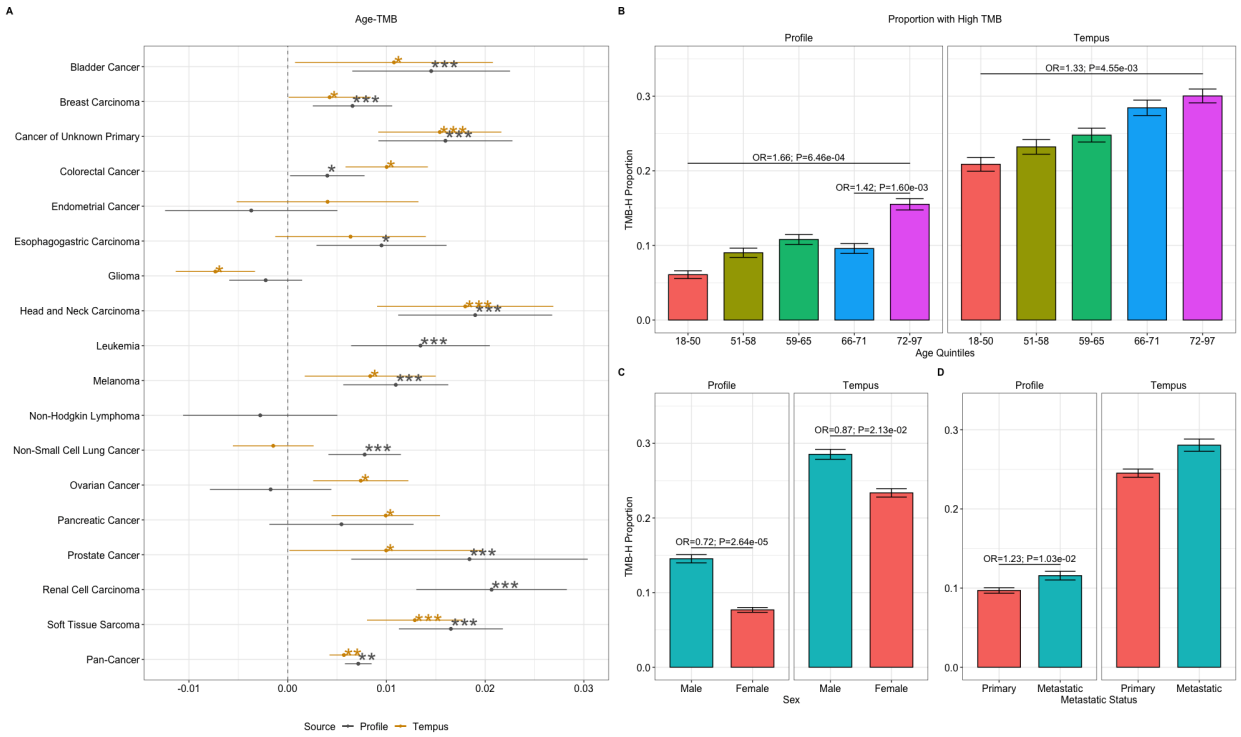


Figure 4.4: **Clinical features are associated with TMB.** (A) Forest plot of the age - TMB beta and the 95% confidence interval for each cancer and pan-cancer meta-analysis. (* - nominal significance; *** - Bonferroni significance; ** - significant meta-analysis) (B) Bar graph indicating the proportion of individuals with TMB-H (TMB $\geq$ 10) pan-cancer by age quintile. Significant odds ratios and their p-values are included. (C) Bar graph showing the proportion of TMB-H patients pan-cancer, stratified by sex with the corresponding significant odds ratio and p-value. (D) Bar graph of proportion of TMB-H split by metastatic status with the significant odds ratio and p-value included.

in the Tempus cohort (as there were sufficient samples for the cancer type and feature), of which 15 were nominally significant (enrichment test $p = 1.39 \times 10^{-13}$) with 9 of the 26 remaining significant after Bonferroni correction.

We observed multiple highly significant associations between age at diagnosis and increased TMB across multiple cancer types in both cohorts. In the pan-cancer meta-analysis, increased age was highly significantly associated with increased TMB in both cohorts (Profile: $\beta = 0.007$, $p = 3.68 \times 10^{-25}$, Tempus: $\beta = 0.006$, $p = 2.43 \times 10^{-14}$; Fig 4.4 sub-figure A). We note that while the effect was extremely significant, the effect size in both cohorts indicated only a small increase in the normalized TMB with each additional year. Within

individual cancers, we observed ten significant associations in Profile. For the 8 out of 10 cancer types that were also present in Tempus, 7 out of 8 of the age - TMB associations were nominally significant, of which three remained significant after Bonferroni correction (Fig 4.4 sub-figure A). Five of the per-cancer age - TMB associations were novel findings without prior evidence in the literature; we tested 4 out of 5 in Tempus and found three were significantly associated (Table 4.1).Overall, older age has long been linked with cancer and more somatic mutations, both pan-cancer and in individual cancers [3, 21, 87]. In fact, of the cancers testable in both cohorts, only three (endometrial cancer, glioma, and pancreatic cancer) did not exhibit at least a nominal positive association between age and TMB.

We observed positive but generally weaker associations between age and CNB in both cohorts. In the pan-cancer meta-analysis, older patients had more somatic CNVs in both cohorts (Profile: $\beta = 0.003$, $p = 4.21 \times 10^{-4}$, Tempus: $\beta = 0.002$, $p = 4.33 \times 10^{-2}$). Additionally, age was significantly associated with CNB in five cancer types in the Profile cohort with the positive associations in glioma and ovarian cancer also present in the Tempus cohort. As CNB based only on deep events has not been previously explored, all five findings were novel. When considering the alternative definition of somatic CNV burden (All CNB) the results were consistent with the primary, deep event-based definition, and have been previously reported (Table 4.1) [86].

Interestingly, we generally observed a protected effect of female sex for TMB and CNB, but had conflicting significant results in the pan-cancer sex - CNB association across the two cohorts. For sex - TMB, the well established protective effect in melanoma was significant in both cohorts (Profile: $\beta = -0.295$, $p = 1.44 \times 10^{-4}$; Tempus: $\beta = -0.374$, $p = 2.77 \times 10^{-4}$) [59, 137]. However, the previously reported pan-cancer protective effect of female sex was not significant in Profile but was significantly associated in Tempus (Profile: $\beta = -0.034$,$p = 8.45 \times 10^{-2}$; Tempus: $\beta = -0.050$, $p = 2.16 \times 10^{-2}$). In the pan-cancer sex - CNB meta-analysis, female sex was significantly associated with lower CNB in the Profile cohort ($\beta = -0.059$, $p = 4.04 \times 10^{-3}$), whereas the association was significant but in the opposite direction in the Tempus cohort ($\beta = 0.111$, $p = 1.76 \times 10^{-6}$); this difference in

direction of effect is likely due to cohort heterogeneity (see Discussion). Within individual cancers, we identified a significant association in esophagogastric cancer between both sex - CNB and sex - All CNB. While the CNB discovery is novel, the All CNB discovery was previously reported (Table 4.1) [86].

Metastatic status was significantly associated with increased CNB and TMB pan-cancer for all three definitions of somatic burden (metastatic - TMB $\beta = 0.059$, $p = 2.55 \times 10^{-3}$; metastatic - CNB $\beta = 0.164$, $p = 7.17 \times 10^{-16}$; metastatic - All CNB $\beta = 0.200$, $p = 1.73 \times 10^{-26}$). While the metastatic - TMB and metastatic - CNB associations were both tested in Tempus, only the metastatic - TMB association was significant in the tumor sample ($\beta = 0.045$, $p = 3.13 \times 10^{-2}$) but not in the normal sample ($p = 7.32 \times 10^{-2}$). We note, however, that in Profile, metastasis is defined based on the biopsy site whereas in Tempus metastasis is defined based on the disease stage even when the primary tumor was sequenced (see Methods). Within cancers, we observed two significant metastatic - TMB associations (breast carcinoma and non-small cell lung cancer), with the association in breast carcinoma being nominally significant in the Tempus cohort (Profile: $\beta = 0.171$, $p = 9.15 \times 10^{-4}$; Tempus: $\beta = 0.119$, $p = 3.44 \times 10^{-2}$). We also observed four cancers with metastatic - CNB associations and six metastatic - All CNB associations. Overall, our findings indicate that metastatic status is correlated with a higher mutational load and that this association was a result of the tumor site itself. This is to say, it is not a function of disease stage because the Tempus indicator for advanced stage/metastatic cancer did not reflect these significant findings [135].

Finally, we conducted additional analyses of the significant age - TMB, sex - TMB, and metastatic - TMB associations using a logistic regression to test if the clinically relevant indicator TMB-H (TMB ≥ 10) was also associated with the clinical feature. We found that pan-cancer, all three clinical features were significantly associated with TMB-H, with the sex and age associations significant in both cohorts (Fig 4.4 sub-figure C and D). For example, we observed 1.66 times the odds of TMB-H for patients 72-98 years of age compared to those aged 18-50 (the two quantile extrema) with a significant increase also observed in

Tempus (Fig 4.4 sub-figure B). Interestingly, while both increased age and male sex are associated with an increased probability of having TMB-H, age was more impactful. An older woman (age 72-97) would have a 12% probability of qualifying for immunotherapy based on the TMB-H threshold in Profile and a 17% probability in Tempus while a young man (age 18-50), would have a 6% and 15% probability, respectively. This is a 55% decrease in Profile and a 17% decrease in Tempus. In addition to the pan-cancer discoveries, we analyzed the seven significant individual cancer age - TMB associations. We observed three cancers with a significant age - TMB-H association in Profile with the melanoma and soft tissue sarcoma associations also significant in Tempus. We also analyzed the sex - TMB association in melanoma and observed female sex was protective for TMB-H in both cohorts (Profile: $OR = 0.52$, $p = 2.96 \times 10^{-4}$; Tempus: $OR = 0.47$, $p = 9.04 \times 10^{-4}$). Overall, our findings indicate that sex and age have a clinically meaningful association with TMB and to a lesser extent CNB across multiple cancer types in both tested cohorts.

4.2.3 Fine-scale European ancestry influences somatic burden

We next explored the association of TMB and CNB with fine-scale genetic ancestry within the European population, focusing on Northwest/Southeast Europe (NW-SE) as a continuous cline and Ashkenazi/non-Ashkenazi Jewish (AJ-non AJ) ancestry as a dichotomous feature (see Methods; Fig 4.5 sub-figure A). We note that while ancestry was estimated based on germline variation, it additionally reflects lifestyle and other non-genetic factors relevant to cancer risk, all of which may influence the accumulation of somatic events. To our knowledge, this is the first examination of fine-scale genetic ancestry on somatic burden.

When considering increased Southeast European ancestry (SE) along the NW-SE cline, we generally observed an increase in TMB in Profile while Tempus generally lacked sufficient power for discovery. In the pan-cancer meta-analysis, we observed increased SE was significantly associated with increased TMB in both cohorts (Profile: $\beta = 0.107$, $p = 5.8 \times 10^{-35}$; Tempus: $\beta = 0.025$, $p = 6.74 \times 10^{-3}$). However, we did not observe a significant effect in the Tempus normal sample likely due to a further decrease in sample size and the nor-

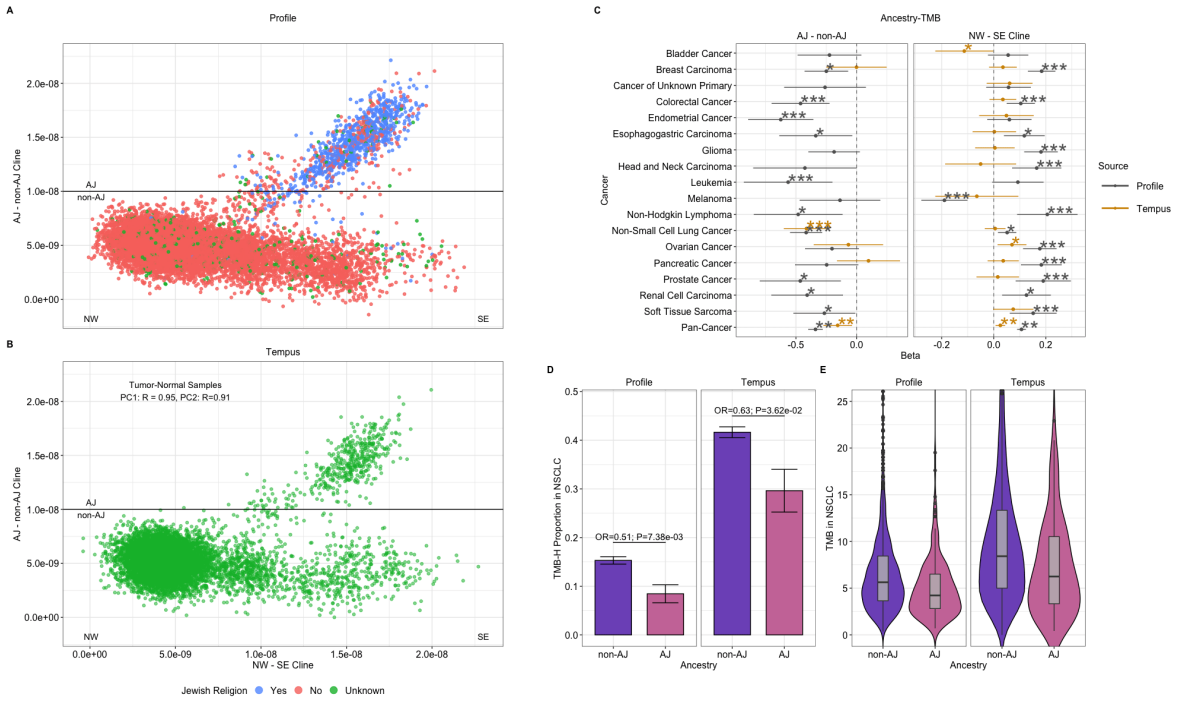


Figure 4.5: **Fine scale ancestry is associated with TMB**. (A) Inferred European ancestry in Profile, color coded by self-reported religion: Jewish religion (red), non-Jewish religion (blue), and unknown religious status (green). The x-axis represents the Northwest-Southeastern cline and the y-axis indicates non-Ashkenazi Jewish versus Ashkenazi Jewish (AJ) ancestry with a vertical line at y = 1.0×10^{-8} indicating the dichotomous variable threshold. (B) Inferred European ancestry in Tempus, with all points shown in green as religion is unknown. The x-axis, y-axis and indicator variable threshold are identical to subfigure (A). (C) Forest plots of the two ancestry-TMB associations with the beta and the 95% confidence interval for each cancer and a pan-cancer meta-analysis for Profile (grey) and Tempus (gold). The left panel shows the AJ indicator results, and the Northwest-Southeastern cline results are in the right panel. (* - nominal significance; *** - Bonferroni significance; ** - significant meta-analysis). (D) Bar graph indicating the proportion of individuals with TMB-H (TMB $\geq$ 10) in non-small cell lung cancer stratified by AJ ancestry with each cohort in a separate panel. Significant odds ratios and their p-values are included. (E) Violin plot of TMB in non-small cell lung cancer with Profile in the left panel and Tempus in the right panel. Each cohort is stratified by AJ ancestry.

mal sample's own sources of noise (e.g. lower coverage). When considering the effect of the NW-SE cline on TMB within individual cancers, there were ten significant associations. While 9 out of 10 cancer types were testable in Tempus, we only saw a nominally significant effect in ovarian cancer ($p = 1.3 \times 10^{-2}$). For 9 out of 10 significant discoveries in Profile and the significant pan-cancer association, we observed an increase in TMB with increasing

SE ancestry. Surprisingly, melanoma was the only cancer showing a significant decrease in TMB with increasing SE ancestry ($\beta = -0.190$, $p = 3.21 \times 10^{-5}$) though this finding was not observed in Tempus. When we considered the phenotype TMB-H for the two cancers significantly associated in both cohorts, we did not observe any significant associations. Lastly, we report a positive pan-cancer association between increased SE ancestry and increased CNB in Profile ($\beta = -0.024$, $p = 6.75 \times 10^{-3}$), but the finding was not significant in Tempus ($p = 1.94 \times 10^{-1}$).

Turning to AJ ancestry, we observed a significant decrease in TMB relative to individuals with non-AJ ancestry. In the pan-cancer meta-analysis, the association with lower TMB was significant in both cohorts (Profile: $\beta = -0.339$, $p = 3.19 \times 10^{-28}$, Tempus: $\beta = -0.156$, $p = 9.84 \times 10^{-3}$) though the smaller normal sample in Tempus was only borderline significant ($p = 5.14 \times 10^{-2}$). When considering individual cancers in Profile, there were four significant associations. Only the significant association in non-small cell lung cancer was testable in the Tempus cohort (due to having > 10 AJ ancestry individuals), and we observed a significant association (Profile: $\beta = -0.416$, $p = 9.77 \times 10^{-10}$; Tempus: $\beta = -0413$, $p = 1.61 \times 10^{-5}$; Fig 4.5). We then followed up on the AJ-non AJ - TMB association using the TMB-H indicator; we observed a significantly lower rate of TMB-H in non-small cell lung cancer in both cohorts for individuals with Ashkenazi Jewish ancestry with comparable effect sizes (Profile: $OR = 0.51$, $p = 7.38 \times 10^{-3}$; Tempus: $OR = 0.63$, $p = 3.62 \times 10^{-2}$; Fig 4.5 sub-figures D and E). There were no significant associations between AJ ancestry and CNB.

While the pan-cancer associations with TMB were observed in both cohorts, generally the per-cancer associations in Profile were not significant in Tempus. We, however, did find that the effect directions were consistent for the significant NW-SE cline discoveries (8/9, $p = 2.00 \times 10^{-2}$); this same test could not be performed for the dichotomous AJ-non AJ variable due to only one testable association in Tempus. To better understand the inconsistency in results, we considered differences between the cohorts. Overall, there were a number of systemic differences (see Discussion and Supplementary Materials; **Supplementary Figures 8-12**), but here we focus on ancestry cline specific differences. In Fig 4.5 sub-figure B,

we observed a much lower density of individuals along the clines in Tempus, with variance along both clines in Tempus approximately half of the variance in Profile (NW-SE cline ratio of variances: 0.64, NW-AJ cline ratio of variances: 0.56). As a result only 4 out of 14 cancers in Tempus met the threshold for inclusion, ten or more individuals with Ashkenazi ancestry, for the AJ-non AJ analyses. While this was not the case with the NW-SE cline as it is continuous, there was a noticeable NW mode. These variance differences in the independent variable coupled with overall differences between the cohorts limits statistical power.

4.2.4 Germline polygenic risk scores are associated with somatic burden

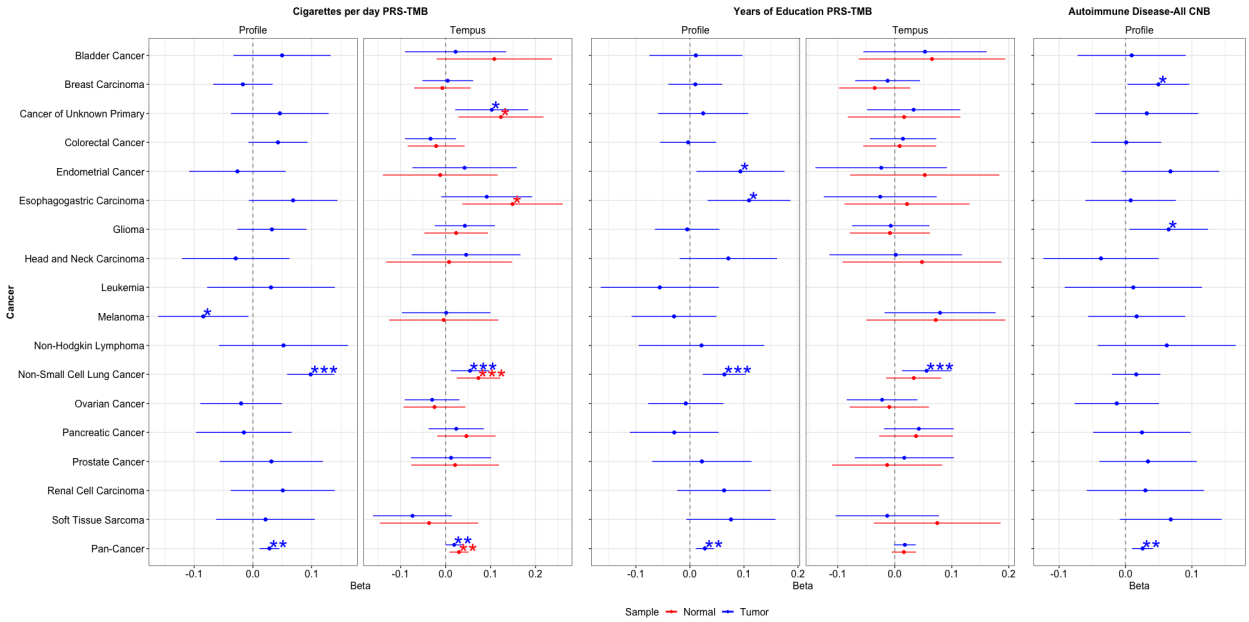


Figure 4.6: **Polygenic risk scores are associated with somatic burden.** Forest plots showing the estimated effect size and the 95% confidence interval in each sub-figure. All sub-figures are stratified by cohort with Profile on the left and Tempus on the right. Tumor samples are in blue and normal samples in red. (* indicates nominal significance; ** shows Bonferroni significance; ** represents a significant meta-analysis) (A) Forest plot of Cigarettes Per Day PRS - TMB associations (B) Forest plot of Years of Education PRS - TMB associations (C) Forest plot of Autoimmune Disease PRS - All CNB associations.

We next examined the germline influence on somatic burden through polygenic risk scores (PRS). There are numerous known cancer risk factors, many of which have a genetic underpinning and some of which directly induce somatic mutations; with this in mind, we

selected 14 relevant phenotypes, including cigarettes per day, autoimmune disease diagnosis, and ease of tanning from publicly available GWAS to identify PRS associated with TMB and CNB [88, 93, 107, 108, 121, 133, 136, 171]. We used a pruning and thresholding approach to construct eight PRS per phenotype and then conducted an association test between each PRS and TMB and CNB in the Profile cohort [166]. For each PRS, the most significant PRS threshold was selected and then brought forward for testing in the Tempus cohort.

We began with an exploration of the associations between PRS and TMB where we identified nine significant discoveries. There were three pan-cancer associations in Profile: smoking (cigarettes per day), educational attainment (years of education), and white blood cell count. Of these, the PRS for cigarettes per day was also significant and positively associated with TMB in the Tempus cohort (Profile: $\beta = 0.028$, $p = 1.06 \times 10^{-3}$; Tempus: $\beta = 0.019$, $p = 4.67 \times 10^{-2}$; Fig 4.6). We tested individual cancers separately and identified six significant PRS - TMB associations of which three were also significant in Tempus: smoking (cigarettes per day) and education attainment (years of education) in non-small cell lung cancer and ease of tanning in melanoma. Of the PRS - TMB findings reported, only the pan-cancer education attainment discovery has previously been reported in TCGA (Table 4.1) [148].

We next sought to estimate the causal effect of cigarettes per day on the number of somatic mutations using a Mendelian Randomization approach with the raw (unnormalized) TMB phenotype (see Methods). Within non-small cell lung cancer, every ten additional cigarettes resulted in almost two additional somatic mutations ($\beta = 1.88$, $p = 7.00 \times 10^{-3}$) while the pan-cancer regression was not significant ($\beta = 0.26$, $p = 4.35 \times 10^{-1}$). In addition, we similarly explored the causal effect of tanning ability and observed that a limited ability to tan (relative to the ability to tan well/moderately well) resulted in over twelve additional somatic mutations ($\beta = 12.68$, $p = 1.07 \times 10^{-3}$).

Another significant discovery was the educational attainment (EA) PRS which was associated with lower TMB both in non-small cell lung cancer and pan-cancer. The pan-cancer association has been previously reported but was not significant in Tempus (Profile:

$\beta = -0.028$, $p = 1.37 \times 10^{-3}$; Tempus: $\beta = -0.018$, $p = 6.99 \times 10^{-2}$, Table 4.1) [148]. The effect in non-small cell lung cancer, however, was significant in both cohorts (Profile: $\beta = -0.065$, $p = 1.35 \times 10^{-3}$; Tempus: $\beta = -0.056$, $p = 1.15 \times 10^{-2}$). In order to determine whether the genetic effect was mediated by a direct measure of EA, we re-evaluated the EA PRS - TMB association in non-small cell lung cancer and included an indicator for graduating college. In this model, the decrease in TMB was significantly associated with graduating college and was no longer associated with the EA PRS. When considering the phenotype TMB-H, the EA PRS association was not significant in Profile ($p = 3.16 \times 10^{-1}$), in fact only the tanning PRS in Melanoma was significantly associated with TMB-H in Profile ($OR = 1.37$, $p = 9.28 \times 10^{-4}$), but it was not significant in Tempus ($p = 2.19 \times 10^{-1}$). However, when we considered the association between graduating college and TMB-H without including the EA PRS, the effect was significant ($OR = 0.708$, $p = 8.19 \times 10^{-3}$). To place this in context, this means that a 66 year old man with a primary tumor (and all other covariates set to the mean value) who did not graduate college has a 15% probability of qualifying for immunotherapy based on the TMB-H threshold while that same man would have a 25% decrease in probability had he graduated from college. While in this scenario the man who did not graduate college was more likely to qualify for immunotherapy, we note that lower socioeconomic status (including lower levels of education) are negatively correlated with cancer prognosis and aggressiveness in cancer treatment [164]. We, therefore, decided to directly test whether graduating college influenced TMB both within individual cancers as well as pan-cancer. We observed a significant effect both pan-cancer ($\beta = -0.084$, $p = 1.20 \times 10^{-5}$) and in two individual cancers (non-small cell lung cancer: $\beta = -0.261$, $p = 1.50 \times 10^{-9}$; cancer of unknown primary: $\beta = 0.263$, $p = 2.96 \times 10^{-3}$). However, when we considered the phenotype TMB-H, only the effect in non-small cell lung cancer (as shown above) was significant.

Turning to CNB, we identified only a single association in Profile, EA PRS in melanoma, which was not significant in the Tempus cohort. For All CNB, there were six significant associations though they could not be tested in Tempus (differing CNB definitions). We

note that all findings are novel and highlight the positive effect size of the autoimmune disease PRS on All CNB for both non-hodgkin's lymphoma and pan-cancer which may indicate interesting biology, since previous work has also linked autoimmune disease and cancer (Fig 4.6) [43, 53]. Finally, we conducted a meta-analysis of three cohorts, the two main cohorts presented here and TCGA (see below and Methods). We concluded with this analysis to explore whether there was evidence of polygenic germline influences via PRS on TMB and CNB that the current sample sizes were ill-powered to identify. In total, we found 19 significant PRS associations ($p < 3.57 \times 10^{-3}$) with two significant associations for PRS - CNB, four for PRS - All CNB, and 13 for PRS - TMB. Of these associations, seven were significant in Profile after Bonferroni correction. The other 12 findings were only discovered via the meta-analysis which implies that further exploration in larger cohorts is warranted.

4.2.5 Germline and somatic variants jointly impact overall survival

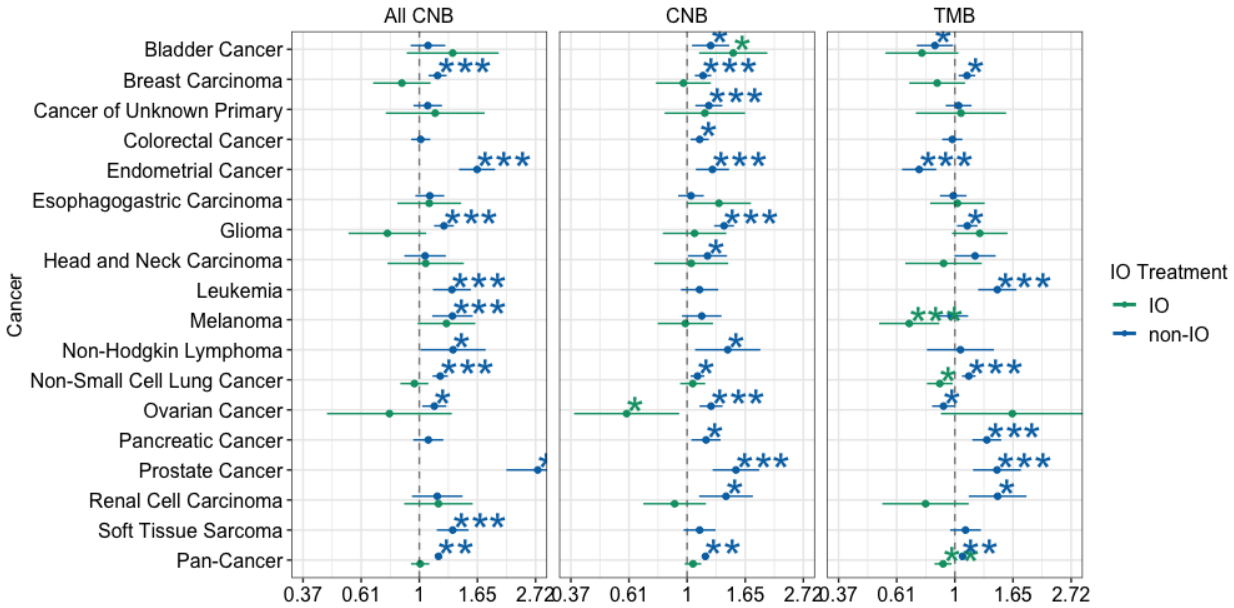


Figure 4.7: **Somatic burden is associated with overall survival.** Forest plots showing the estimated effect size and the 95% confidence interval of somatic burden both immunotherapy (IO) patients and non-IO patients. Each sub-figure corresponds to a different somatic burden definition, and within each panel, IO patients are blue and non-IO patients are green. (A) All CNB - OS association (B) CNB - OS association (C) TMB - OS association.

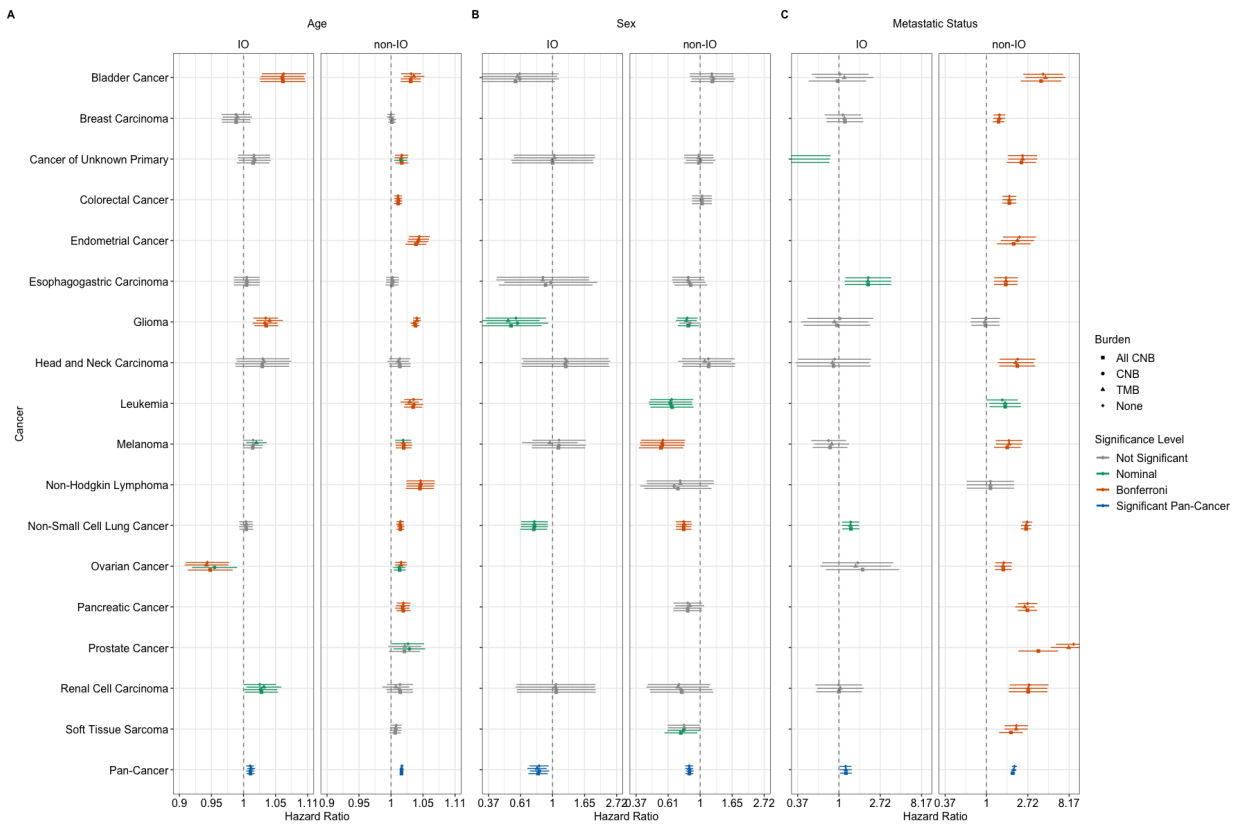


Figure 4.8: **Clinical features are associated with overall survival.** Forest plots showing the estimated effect size and the 95% confidence interval of clinical features for both immunotherapy (IO) patients and non-IO patients. Each sub-figure corresponds to a different clinical feature, and within each sub-figures there is a panel for IO patients and one for non-IO patients. We condition on various somatic burden types and use a unique symbol for each definition and use color to indicate the significance of the regression (A) Age - OS association (B) Sex - OS association (C) Metastatic status - OS association.

Previous work has linked both TMB and CNB as well as clinical features with cancer outcomes [30, 34, 66, 95, 129, 132, 142, 152, 167]. Here, we sought to further investigate these effects in a large cohort as well as explore the effect of polygenic germline features on overall survival (OS). We restrict our analyses to the features significantly associated with TMB and CNB as reported above to reduce the multiple testing correction. We first analyzed the direct effect of these features on OS and if significant, tested whether the effect on OS was mediated by TMB or CNB. We also tested all features for whether the features and somatic burden had an interaction effect on OS. These analyses were restricted to 11,973 patients from the Profile cohort who had treatment and survival measurements

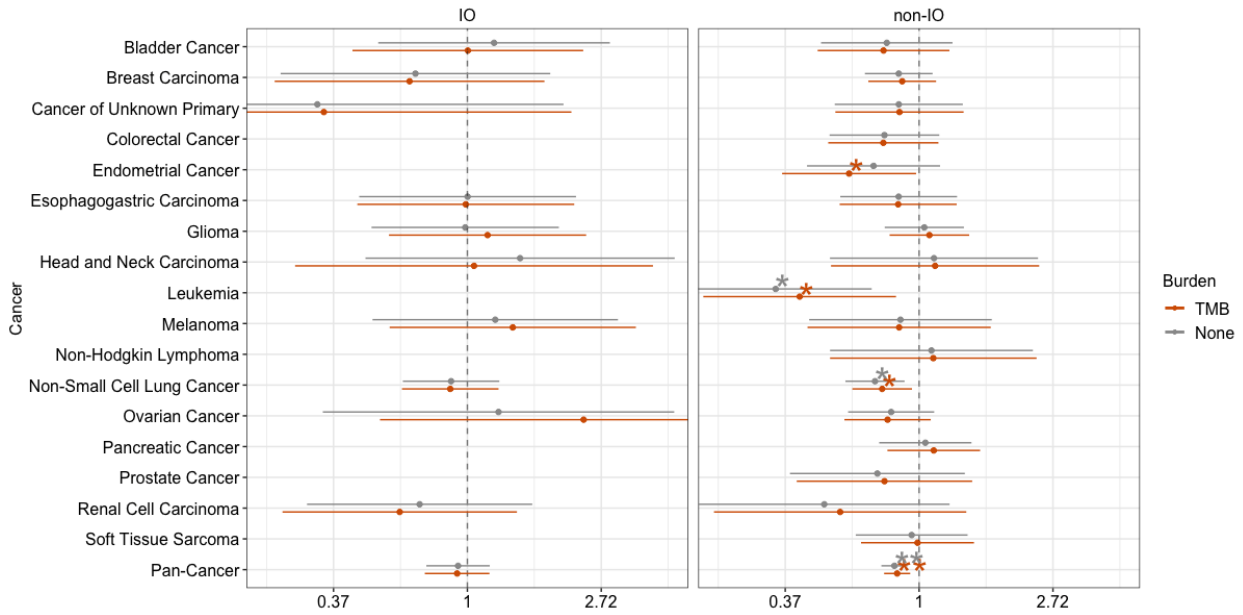


Figure 4.9: **Ashkenazi Jewish ancestry is associated with overall survival.** Forest plots showing the effect size and 95% confidence interval of Ashkenazi Jewish ancestry for both immunotherapy (IO) patients and non-IO patients. We include the estimate with and without conditioning on tumor mutational burden (TMB).

readily available. These patients were treated at Dana-Farber Cancer Institute and 1,415 of these patients were immunotherapy recipients (IO) while the remaining 10,558 patients were immunotherapy naive (non-IO).

We first explored the influence of TMB and CNB on OS which both confirmed previous findings in a large cohort as well as explored additional cancers. When we considered only IO recipients, we observed a pan-cancer protective effect of TMB on OS ($HR = 0.903$, $p = 6.61 \times 10^{-3}$) while there was no relationship between CNB and OS or All CNB and OS. When considering individual cancers only melanoma had a significant protective effect ($HR = 0.675$; $p = 2.94 \times 10^{-3}$) though the effect in non-small cell lung cancer was nominally significant ($HR = 0.878$, $p = 2.19 \times 10^{-2}$). We then considered the effect of TMB-H on OS where we observed an even stronger pan-cancer protective effect for IO recipients ($HR = 0.650$, $p = 2.99 \times 10^{-4}$) as well as a protective effect of in melanoma ($HR = 0.585$, $p = 2.92 \times 10^{-2}$) and non-small cell lung cancer ($HR = 0.626$, $p = 4.76 \times 10^{-3}$). We then analyzed the non-IO patients and saw that both TMB and CNB were associated with

increased risk (TMB - OS: $HR = 1.068$, $p = 1.52 \times 10^{-6}$; CNB - OS: $HR = 1.170$, $p = 5.19 \times 10^{-31}$; All CNB - OS: $HR = 1.180$, $p = 1.63 \times 10^{-30}$; Fig 4.7). When considering individual cancers, there were 19 significant associations: five TMB - OS, six CNB - OS, and eight All CNB - OS. The vast majority of these significant associations aligned with the pan-cancer results and showed somatic burden conferred increased risk. The one exception was in endometrial cancer which indicated increased TMB had a protective effect on OS ($HR = 0.736$, $p = 4.27 \times 10^{-5}$). When we considered the phenotype TMB-H, we did not observe a pan-cancer effect but did see TMB-H was protective in endometrial cancer ($HR = 0.374$, $p = 2.7510^{-3}$).

We then turned to the clinical features significantly associated with TMB or CNB in order to determine how they influence OS as well as whether their effects are mediated by TMB or CNB. We began with a positive control and saw that consistent with prior work, age, male sex and metastatic status were associated with poorer survival within multiple individual cancers as well as pan-cancer [30, 95, 129, 149, 152]. This is true for both IO recipients (age - OS: $HR = 1.010$, $p = 1.21 \times 10^{-3}$, sex - OS: $HR = 0.810$, $p = 5.46 \times 10^{-3}$; metastatic - OS: $HR = 1.173$, $p = 3.09 \times 10^{-2}$; Fig 4.8) and non-IO patients (age - OS: $HR = 1.017$, $p = 7.50 \times 10^{-49}$; sex - OS: $HR = 0.843$, $p = 8.23 \times 10^{-8}$; metastatic - OS: $HR = 1.969$, $p = 2.17 \times 10^{-109}$; Fig 4.8). The one exception was ovarian cancer, where increased age was significantly protective for IO recipients ($HR = 0.945$, $p = 1.08 \times 10^{-3}$). We then tested these associations conditioned on the effect of TMB and CNB separately and did not observe a mediating effect indicating that while increased age, male sex and metastatic tumor sites were associated with increased somatic burden, these clinical features have an association with OS independent of their relationship with TMB or CNB. In fact, when we consider the age - OS association amongst IO recipients, we saw the effect size was consistent but the association signal was more significant (age - OS (no somatic burden covariate): $HR = 1.010$, $p = 1.21 \times 10^{-3}$; controlling for All CNB: $HR = 1.011$, $p = 6.62 \times 10^{-4}$; controlling for CNB: $HR = 1.011$, $p = 9.26 \times 10^{-4}$; controlling for TMB: $HR = 1.012$, $p = 2.26 \times 10^{-4}$). Lastly, in addition to controlling for somatic burden, we tested a number of other sources

of confounding on the protective effect of age in ovarian cancer amongst IO recipients. We saw it remained after controlling for the cancer subtype ($HR = 0.957$, $p = 3.0 \times 10^{-2}$). It was also still significant after accounting for biases from the course of care including: line of treatment, concurrent treatment with chemotherapy, and whether the patient was sequenced after treatment began ($HR = 0.925$, $p = 2.5x10^{-4}$).

Finally, we considered how the polygenic germline features: PRS and fine-scale ancestry impacted survival, beginning with IO recipients. We first analyzed their marginal effect on OS where we observed no significant associations; therefore, we did not test their effect conditioned on the effect of TMB or CNB. We next tested whether either TMB or CNB and the polygenic germline features had an interaction effect on OS where we identified two significant associations. For patients with non-small cell lung cancer, there was an interaction between TMB and the EA PRS ($HR = 0.859$, $p = 1.13 \times 10^{-2}$). This indicates that patients with a higher EA PRS and higher TMB fared better though the EA PRS did not have a significant marginal effect ($p = 7.77 \times 10{-}1$) and TMB itself had a protective effect ($HR = 0.836$, $p = 3.02 \times 10^{-3}$). We then included an indicator for graduating college in the regression where we observed that neither the marginal effects of TMB and EA PRS nor the interaction effect between TMB and EA PRS were significantly associated with OS (TMB: $p = 1.70 \times 10^{-1}$; EA PRS: $p = 9.44 \times 10^{-1}$; interaction effect: $p = 3.38 \times 10^{-1}$) while the indicator for college was significantly associated ($HR = 0.656$, $p = 7.42 \times 10^{-3}$). The other significant interaction effect in IO patients was observed in melanoma where there was a protective interaction between higher All CNB and increased polygenic risk of developing lung cancer amongst smokers ($HR = 0.718$, $p = 8.74 \times 10^{-3}$). While neither All CNB nor the PRS had a marginal effect on OS, this interaction indicates patients with higher somatic CNV burden and a higher PRS fare better than the baseline. This means that while CNB correlates with immune evasion which itself is linked with cancer outcomes, there may also be underlying germline factors interacting with CNB which has a protective influence on survival [34, 154]. In order to determine whether the risk was related to the genetic risk of smoking, we re-analyzed the association but included the cigarettes per day

PRS as an independent variable. We observed no mediating effect of the genetic liability of cigarettes per day ($p = 4.34 \times 10^{-1}$; interaction between All CNB and PRS: $HR = 0.730$, $p = 1.30 \times 10^{-2}$) which implies the interaction is not due to the genetic risk of smoking cigarettes directly but rather the genetic risk of developing lung cancer.

We now conclude with the PRS and fine-scale ancestry associations with survival in non-IO patients. When considering their marginal effect, we observed a significant protective effect of Ashkenazi Jewish (AJ) ancestry in two individual cancers as well as pan-cancer (non-small cell lung cancer: $HR = 0.719$, $p = 3.47 \times 10^{-3}$; leukemia: $HR = 0.343$, $p = 3.40 \times 10^{-3}$; pan-cancer: $HR = 0.831$, $p = 1.93 \times 10^{-4}$;Fig 4.9). We then tested whether these associations were mediated by TMB but found all three effects of AJ ancestry remained significant (non-small cell lung cancer: $p = 1.52 \times 10^{-2}$; leukemia: $p = 1.51 \times 10^{-2}$; pan-cancer: $p = 9.89 \times 10^{-4}$). We also observed that increased Southeastern (SE) European ancestry was significantly protective in head and neck carcinoma ($HR = 0.634$, $p = 1.04 \times 10^{-3}$) and that this effect remained significant after controlling for TMB ($p = 6.09 \times 10^{-4}$). We then tested for significant interaction effects between TMB or CNB and polygenic germline features and observed one significant effect in non-IO patients. In pancreatic cancer, there was a protective interaction between increased SE ancestry and increased TMB ($HR = 0.860$, $p = 4.30 \times 10^{-2}$). This interaction indicates that while alone SE ancestry is protective and increased TMB is associated with poorer outcomes, when a patient has more SE ancestry and higher TMB, some of the negative effect of high TMB is mitigated.

4.2.6 Comparison of our findings to discoveries reported in TCGA

We next examined the cross replication of the 26 novel TMB and CNB associations testable in The Cancer Genome Atlas (TCGA) as well as 24 previously reported findings, 12 of the 24 were significant discoveries in Profile (Table 4.1). As our analysis pipeline differs from previous work, we re-analyzed the previous discoveries. In particular, our analysis pipeline, unlike previous work is restricted to patients with European ancestry and who were microsatellite stable to prevent confounding by continental ancestry and race based

Figure 4.10: **Genetic ancestry in TCGA.** (A) Inferred continental ancestry in TCGA, color coded by self-reported race. We restrict the analyses to individuals within two standard deviations of the mean inferred ancestry of self-reported white individuals with the boundaries shown by the black rectangle. (B) Zoomed in plot of the black rectangle in sub-figure A of TCGA. (C) Inferred European ancestry in TCGA, color coded by self-reported religion: Jewish religion (red), non-Jewish religion (blue), and unknown religious status (green). The x-axis represents the Northwest-Southeastern cline and the y-axis indicates non-Ashkenazi Jewish versus Ashkenazi Jewish ancestry.

biases from inequitable access to care as well as spurious associated due to hypermutability. We were able to replicate 14 out of 24 findings using our pipeline, 9 of which were among the 12 significant Profile findings. Of the 27 novel discoveries, we were able to replicate 11 nominally with two events being pan-cancer associations, and 5 of the 26 remained significant after Bonferroni correction. We note that metastatic status, fine-scale ancestry and survival analyses were not considered. We excluded metastatic status and survival analyses due to the limited availability of metastatic tumors and survival measurements in TCGA, and we forgo replicating ancestry discoveries due to TCGA facing the same limitations as Tempus while also having less than half of the sample size (Fig 4.10).

4.3 Discussion

In this chapter, we identified numerous novel significant associations between both clinical and polygenic germline features and both CNB and TMB as well as replicated many results from previous studies. We observed that age and sex have a strong effect on TMB and a weaker effect on CNB across many contexts. While metastatic sites resulted in increased TMB and CNB, this was in a very context specific way and was directly related to the site sequenced instead of metastases as a function of disease stage as defined in Tempus. Additionally, we found that fine-scale genetic ancestry was associated with TMB pan-cancer, both through the Northwest-Southeast European cline and Ashkenazi Jewish ancestry, and these novel findings warrant further exploration. We additionally observed that smoking causally increased TMB by approximately one mutation per pack per day and that the limited ability to tan causally increased TMB by more than 12 mutations. We were also able to link many of these significantly associated features with overall survival (OS) and found that while these features were significantly correlated with somatic burden, their effect on OS was not mediated by TMB or CNB. In addition to the marginal associations, we observed three interaction effects between germline genetics and somatic burden on OS. These findings indicate that not only is there a polygenic germline influence on somatic burden, somatic and germline genetics jointly impact survival.

Our study has multiple limitations. First, heterogeneity between the Profile and Tempus cohorts resulted in some findings that were specific to one cohort and may not generalize. To understand what is contributing to cohort heterogeneity, we first considered the distribution of clinical features across cohorts both pan-cancer and for each individual cancer (Figs 4.11-4.15). When considering the distribution of age and the proportion of female patients, we observed consistency both pan-cancer and within cancers. When considering metastatic status, we observe the pan-cancer proportion of metastasis was nearly identical in the two cohorts while the proportions within cancers was more variable. Setting aside cancer of unknown primary, some of the difference in metastatic status between cohorts may be due to sampling bias; however, it is also influenced by how we defined metastatic status

in the Tempus cohort. This mislabeling of advanced stage primary sites as metastatic would introduce noise and thereby reduce generalizability. We also considered the difference between the distributions of tumor purity which impacts both somatic and germline variants. Low tumor purity indicates a higher proportion of normal cells in the sample. While this has previously been shown to have a negligible impact on germline imputation, it makes somatic variant calling more difficult [60, 80]. While Tempus has a lower tumor purity reported, we note that the availability of normal-match sampling (78% of samples) mitigates the influence on somatic calling (Supplementary Figure 2) [10].

When considering additional differences in somatic calling, there was discrepancy in how somatic events were defined. This is due, in part, to the fact that precise quantification of TMB and CNB remains an open area of research. While the definitions of TMB were likely comparable phenotypes, CNB was less consistent between cohorts which would therefore impact consistency in results with the degree of impact depending on the power to detect that particular effect size (Supplementary Figures 3-5). While our study shows which findings are generalizable over different calling strategies, consistent phenotype definitions would enable increased power to discover associations across multiple cohorts.

In addition to the limitations within the study design, there were a number of broader considerations. We restricted our analyses to European samples due to sample size, so the generalizability of our findings to cohorts with different ancestral backgrounds is uncertain. Further work is needed to not only determine what is consistent across genetic ancestries, but to explore what is distinct, especially as continental ancestry has been linked with molecular differences within cancers and tissues [20]. Another concern is that patients were primarily sequenced while receiving treatments that can influence somatic burden, but granular treatment history was largely unavailable in Profile and not at all in Tempus. It is possible that some of the cohort heterogeneity between Tempus and Profile in addition to the differences listed above, may also relate to the timing of treatment relative to sequencing as well as treatment type. This possibility is especially likely due to patients in Profile belonging to one institution while Tempus patients originate from multiple institutions which may result

in more heterogeneity in the course of care. Unfortunately, without detailed treatment history, we are unable to quantify these effects. Another broader concern for immunotherapy recipients, is that our analyses cannot rule out reverse causation where patients with TMB < 10 were approved for immunotherapy due to other mitigating circumstances, such as very advanced stage disease or progression while receiving other therapies. In this case the impact of TMB-H on OS may be confounded by systemic differences between who is eligible for immunotherapy instead of the direct effect of TMB-H. Lastly, we only explored overall survival and did not consider treatment response or progression free survival as these measurements were not available.

While we were not the first to explore many of these associations, the vast majority of previous findings leveraged The Cancer Genome Atlas (TCGA), a moderately large, publicly available cohort. Unfortunately, TCGA, which is now over a decade old, has a number of known technical artifacts and biases [16, 28, 81, 127]. This, coupled with it being approximately half of the size of both of the main cohorts and the inherent heterogeneity in cancer cohorts led us to restricting the scope of our replication efforts in TCGA. Fortunately, the ability to generate cohorts containing germline and somatic calling via off-target imputation will result in additional "immunotherapy era" cohorts being generated to aid our understanding of determinants of somatic burden.

With that, the discoveries presented here further uncover host level determinants to the otherwise largely stochastic process of accumulating somatic variants. By understanding the influence of host level features on the somatic landscape of tumors and their joint impact on overall survival, we can move towards personalized oncology that has the ability to treat based on the patient, their tumor, and the interaction between the two. While these findings highlight this interplay, further work is essentially to understand the clinical implications. While previous work has indicated that the TMB-H threshold (TMB ≥ 10) may not be optimal for all cancers, our work indicates that host level factors may also be relevant for setting the TMB-H threshold [106, 146]. Clear statements regarding this, however, necessitate follow-up studies and clinical trials.

In future work, we hope to address the limitations stated above as well as explore new phenotypes. Here, we present associations with somatic burden defined genome-wide, but it is possible that gene level events or hotspot mutations are also genetically determined separately from somatic variant burden. Additionally, clonal and subclonal somatic burden may be influenced differently by clinical and germline genetics which warrants consideration. Lastly, we hope to consider the effect of individual germline variants through genome-wide association studies on a number of somatic variant phenotypes.

4.4 Materials and Methods

4.4.1 Description of the Profile Cohort

Patients receiving routine treatment at Dana-Farber Cancer Institute may consent to participate in the Profile prospective clinical sequencing effort. Each consented tumor biopsy is assayed on one of three panel versions of the targeted capture platform (OncoPanel). The three panel versions target 275, 300, and 447 genes, respectively, and all samples must minimally have 30X coverage for 80% of targets. A clinical bioinformatics pipeline calls all somatic variation, such as single nucleotide variants and copy number variation which are then reviewed by Brigham & Women's Hospital pathologists [52]. We performed germline imputation across all samples using STITCH imputation software as previously described [33, 60]. Germline variants were restricted to those with imputation INFO > 0.4 and minor allele frequency > 0.01. Continental ancestry was computed using imputed dosages and the PLINK2 '–score' function by projecting each sample into the reference PC space generated by SNPweights tools in HapMap populations of European, West African (Yoruban) and East Asian (Chinese) ancestry (Supplementary Figure 1 sub-figure A) [22, 125]. We calculated the mean and standard deviation of both PCs for self-reported white individuals and retain all individuals within two standard deviations of the mean (PC 1: $1.58 \times 10^{-8}(+/- 1.05 \times 10^{-9})$, PC2: $-8.14 \times 10^{-9}(+/- 2.32 \times 10^{-9})$). We then further restricted to 17 cancers with 300 or more microsatellite stable (MSS) patients for a total of $13,131$ individuals (Figure 1 sub-figure

C, Supplementary Figure 1 sub-figure C). Samples were selected and sequenced from patients who were consented under institutional review board (IRB) approved protocol 11-104 from the Dana-Farber/Partners Cancer Care Office for the Protection of Research Subjects. Written informed consent was obtained from participants prior to inclusion in this study. Secondary analyses of previously collected data were performed with approval from the Dana-Farber IRB (DFCI IRB protocol 19-033 and 19-025; waiver of HIPAA authorization approved for both protocols).

4.4.2 Description of the Tempus Cohort

A second independent cohort was generated using a representative population selected from the Tempus genomic database. Each sample was sequenced on one of the three panel versions of the targeted Tempus xT next-generation sequencing platform which respectively target the exons of 595, 596, and 648 genes [10]. The germline imputation and continental ancestry projections were performed in an identical manner as those described above (Supplementary Figure 1 sub-figure B). We again calculated the mean and standard deviation of both PCs for self-reported white individuals and retain individuals within two standard deviations of the mean (PC 1 (tumor): $1.57 \times 10^{-8}(+/- 1.32 \times 10^{-9})$, PC 1 (normal): $1.54 \times 10^{-8}(+/- 1.09 \times 10^{-9})$; PC 2 (tumor): $-9.07 \times 10^{-9}(+/- 2.88 \times 10^{-9})$, PC 2 (normal): $-9.77 \times 10^{-9}(+/- 2.35 \times 10^{-9})$). We then further restrict to the cancers analyzed in Profile resulting in 14 cancers with 200 or more microsatellite stable patients. In total, we have a curated cohort of $10,294$ individuals, the majority of whom (78%) also had a corresponding normal tissue sample (Figure 1 sub-figure C, Supplementary Figure 1 sub-figure D).

4.4.3 Description of the TCGA Cohort

The Cancer Genome Atlas (TCGA) is a well studied, publicly available cohort which has thousands of individuals sequenced both on a germline assay and using whole exome sequencing. We implemented the analysis pipeline used in Profile and Tempus as described above to compare our results to those previously published. The samples were imputed from the geno-

typing array using the Michigan imputation server with the Haplotype Reference Consortium reference panel [105]. Once imputed, we calculated the mean and standard deviation of both PCs for self-reported white individuals and retained individuals within two standard deviations of the mean (PC 1: $1.99 \times 10^{-8} (+/- 1.09 \times 10^{-9})$, PC2: $-7.93 \times 10^{-9} (+/- 2.23 \times 10^{-9})$. We determined microsatellite stability using the publicly reported MSIsensor score and retained individuals with a score < 4 [37, 114]. We also used the consensus tumor purity previously published [5]. We restrict to 11 cancers with 200 or more patients.

4.4.4 Somatic variant calling and outcome generation

The three outcomes assessed in this study are tumor mutational burden (TMB) which is the enumeration of somatic single nucleotide variants (SNVs) and two copy number burden definitions which enumerate somatic copy number variation (CNVs). For TMB, we restrict somatic SNVs called on the coding region of each gene and show the distribution is comparable between cohorts (Figure 1 sub-figure B). We additionally generate the phenotype TMB-H which is an indicator for high TMB (TMB ≥ 10). When conducting logistic regressions using TMB-H as the phenotype, we only consider cancers with > 10 individuals with TMB-H. For copy number burden, one definition considers only deep gains or losses while the other considers all CNVs. For Profile and TCGA, each CNV call indicates whether the alteration is deep or shallow; therefore, we generate the two outcomes using this information. The Tempus cohort did not provide an equivalent indicator of CNV depth. Instead a CNV gain was defined as 8 or more copies detected in 4 consecutive regions or at least 20% of the gene regions while a CNV loss was defined as 0 copies detected in 4 consecutive regions or at least 20% of the gene regions; this is most comparable to the deep CNV calls in Profile (Figure 1 sub-figure B, Supplementary Figure 3-5). For simplicity we will refer to the deep gains and losses definition solely as CNB and the other definition of copy number burden which enumerates all gains and losses as "All CNB". We note that the two definitions are truly distinct with a correlation of 0.32 in Profile and 0.36 in TCGA with only a negligible change in the correlation after quantile normalization.

4.4.5 Association between clinical features and somatic burden

The role of clinical features on somatic burden were tested using a multivariate linear regression in each respective cancer as well as via a pan-cancer meta-analysis. The independent variables of interest were age, sex, and metastatic status and the outcomes TMB, CNB, and All CNB were considered separately. In addition to the features of interest, the model included panel version, tumor purity, the first 5 in-sample PCs as covariates, and in Tempus we also included an indicator for whether the tumor sample had a normal-matching sample. We note that in Profile metastatic status is an indicator variable based on the tumor site (local recurrent or primary site versus metastatic site) while Tempus approximates this feature by an indicator variable for whether the cancer description includes "metastatic". The 5 in-sample PCs were generated using PLINK2 and are described in more detail below [125]. We note that we included the metastatic status indicator in cancer of unknown primary, but we did not treat this indicator as a feature for discovery. It was included in the model to account for any noise caused by primary labeling (e.g. different pathologist).

4.4.6 Association between fine-scale ancestry and somatic burden

We used an external reference panel designed to capture the principle components of the within Europe population structure particularly distinguishing between Northwestern Europe (NW), Southeastern Europe (SE), and Ashkenazi Jewish (AJ) ancestral populations [22]. We project each sample into the corresponding PC space using PLINK2 '–score' [125]. As a sanity check, we confirmed that AJ-non AJ was significantly associated with self-reported Jewish religion ($\rho = 0.69$, $p < 2.2 \times 10^{-16}$; Figure 3 sub-figure A), acknowledging that these are not expected to be perfect surrogates. We convert the AJ - non AJ cline to an indicator with Ashkenazi Jewish ancestry corresponding to a PC value $\geq 1 \times 10^{-8}$ in both cohorts. We only considered cancers for testing (and replication) if there were at least 10 individuals with Ashkenazi Jewish ancestry; while all cancers in Profile met this threshold, only 4 in Tempus did. We separately test the effect of NW-SE cline and the AJ-non AJ indicator and exclude AJ individuals from the NW-SE cline regression. We used a linear regression that

controlled for the effects of sex, age, metastatic status, panel version, and tumor purity, and a covariate for whether the patient has a normal-match was also included in Tempus.

4.4.7 Association between polygenic risk scores and somatic burden

In addition to the initial restriction on imputed SNPs (MAF > 0.01 and INFO > 0.4), we further restrict the germline variants to HapMap3 SNPs and then LD-prune this set using PLINK2 '–indep-pairwise 500kb 0.5' [125]. These independent SNPs are used to calculate the in-sample principal components (PCs) using PLINK2 '–pca approx' and polygenic risk score (PRS) using PLINK2 '–score'. To generate the PRS, we first chose 14 cancer related outcomes from a number of large GWAS studies [88, 93, 107, 108, 121, 133, 136, 171]. We set eight p-value thresholds in the original GWAS beginning with all SNPs and ending with SNPs with a p-value $< 5 \times 10^{-7}$ considering each order of magnitude between (i.e. 5×10^{-X} for $X \in \{0, 1, \ldots, 7\}$). At each threshold, we use the intersection between the LD pruned SNPs and retained discovery GWAS SNPs to generate the projection. Finally, we generated a centered genetic relatedness matrix (GRM) between individuals using GEMMA and performed a linear mixed model between the somatic burden outcomes and each PRS controlling for age, sex, metastatic status, tumor purity, panel version, the GRM, and whether the sample has a normal-matching sample (in Tempus) and in each analysis. In order to reduce multiple testing correction, we selected the threshold with the most significant association between the PRS and somatic outcome per cancer and separately chose the threshold pan-cancer for each PRS. Using this refined list of PRS, we use a Bonferroni corrected for the number of PRS (14) for each cancer and outcome pair. After identifying a significantly associated PRS, the PRS at the same threshold was tested in Tempus, correcting for the number of significantly associated PRS.

We analyzed two significant PRS associations using a Mendelian Randomization approach. For the cigarettes per day PRS, the smoking phenotype is defined in terms of bins of smoking frequency with the difference in means between neighboring bins corresponding to approximately ten cigarettes per day. For the tanning ability PRS, the tanning pheno-

type is a binary variable splitting individuals who tan very/moderately easily from those who mildly/never tan (i.e. burn). For both PRS, we used the most restrictive threshold 5×10^{-7} and retained the peak SNP in each megabase region. We then conducted a genome-wide association study on the untransformed TMB using a linear mixed model controlling for age, sex, metastatic status, tumor purity, panel version and the first 5 in-sample PCs. We retained the set of SNPs that intersected with the peak SNPs for each PRS. Finally, we separately regressed the original cigarettes per day GWAS betas and the original tanning ability GWAS betas onto the TMB GWAS betas.

Survival analyses using cox-proportional hazard model

After conducting associations between the numerous independent features (e.g. age, ancestry, PRS) and both TMB and CNB, we performed follow-up survival analyses on the significant associations using a cox-proportional hazard model. All models regardless of feature of interest include the covariates: age, sex, metastatic status, panel version and tumor purity; the first 5 PCs were also included except when we analyzed how ancestry impacted overall survival (OS). We separately analyze patients who received immunotherapy (IO) and those who did not (non-IO). We began by analyzing the effect of TMB and CNB on OS. We then separately analyzed the impact of clinical features, fine-scale ancestry, and PRS on OS. For the significant associations, we then conditioned on the effect of TMB or CNB and chose the somatic burden based on which was significantly correlated with the feature of interest. Lastly, we conducted interaction analyses between the somatic burden features (TMB and CNB) and germline genetic features (fine-scale ancestry and PRS).

4.4.8 Figures depicting cohort heterogeneity

We considered the distribution of the clinical features: age, sex, metastatic status and tumor purity between Tempus and Profile. We explored these features both pan-cancer and within each individual cancer and depict their similarities and differences in Figs 4.11-4.15 and provide a detailed discussion in section 4.3.

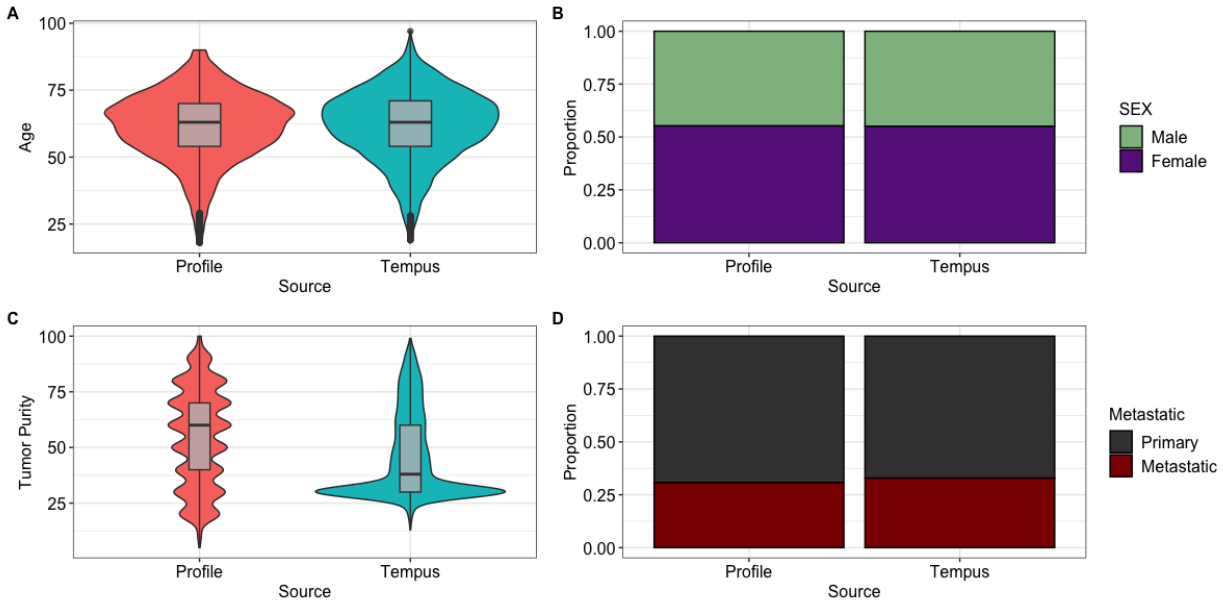


Figure 4.11: **Pan-cancer comparison of covariates.** (A) Violin plot with a box-plot overlaid depicting the distribution of age across cancers in Profile (red) and Tempus (blue) (B) Bar graph of the pan-cancer distribution of sex with purple showing the proportion of women and green the proportion men in each cohort (C) Violin plot with a box-plot overlaid depicting the distribution of tumor purity across cancers in Profile (red) and Tempus (blue) (D) Bar graph of the pan-cancer distribution of metastatic status with red showing the proportion of metastatic cancers and grey the proportion of non-metastatic cancers in each cohort

Cancer	Burden	Independent Variable	Profile		TCGA		Tempus (Tumor)		Tempus (Normal)	
			Beta	P-value	Beta	P-value	Beta	P-value	Beta	P-value
Breast Carcinoma	All CNB	Age	-0.007	1.97E-04	-0.005	1.23E-01	—	—	—	—
Endometrial Cancer	All CNB	Age	0.017	7.87E-05	0.024	1.87E-06	—	—	—	—
Esophagogastric Carcinoma	All CNB	Age	0.006	4.37E-02	0.009	8.24E-02	—	—	—	—
Glioma	All CNB	Age	0.010	1.60E-07	0.022	1.49E-17	—	—	—	—
Non-Small Cell Lung Cancer	All CNB	Age	0.002	3.74E-01	-0.009	2.91E-02	—	—	—	—
Ovarian Cancer	All CNB	Age	0.018	9.85E-10	0.016	2.05E-03	—	—	—	—
Prostate Cancer	All CNB	Age	0.013	1.09E-02	0.021	1.52E-02	—	—	—	—
Soft Tissue Sarcoma	All CNB	Age	0.009	6.37E-04	—	—	—	—	—	—
Pan-Cancer	All CNB	Age	0.004	4.68E-08	0.006	4.56E-08	—	—	—	—
Esophagogastric Carcinoma	All CNB	Sex	-0.350	2.02E-05	-0.346	5.87E-03	—	—	—	—
Head and Neck Carcinoma	All CNB	Sex	-0.122	2.11E-01	-0.047	6.86E-01	—	—	—	—
Renal Cell Carcinoma	All CNB	Sex	-0.085	3.86E-01	-0.412	1.61E-03	—	—	—	—
Soft Tissue Sarcoma	All CNB	Sex	0.036	6.45E-01	—	—	—	—	—	—
Pan-Cancer	All CNB	Sex	-0.051	7.59E-03	-0.042	2.44E-01	—	—	—	—
Breast Carcinoma	CNB	Age	-0.010	7.11E-07	-0.005	1.23E-01	-0.003	1.89E-01	-0.003	2.90E-01
Endometrial Cancer	CNB	Age	0.015	9.96E-04	0.012	2.56E-02	-0.005	2.72E-01	-0.003	6.03E-01
Glioma	CNB	Age	0.015	1.08E-14	-0.004	1.62E-01	0.015	8.07E-12	0.015	1.26E-09
Ovarian Cancer	CNB	Age	0.013	6.47E-05	-0.001	7.96E-01	0.010	1.36E-04	0.009	1.41E-03
Pancreatic Cancer	CNB	Age	0.013	8.48E-04	—	—	-0.003	2.47E-01	-0.004	1.71E-01
Pan-Cancer	CNB	Age	0.003	4.21E-04	-0.002	1.83E-01	0.002	4.33E-02	0.002	4.56E-02
Esophagogastric Carcinoma	CNB	Sex	-0.367	7.28E-05	-0.458	2.06E-04	0.076	5.29E-01	0.119	4.10E-01
Pan-Cancer	CNB	Sex	-0.059	4.04E-03	-0.080	3.25E-02	0.111	1.76E-06	0.113	1.62E-05
Bladder Cancer	TMB	Age	0.015	3.96E-04	0.002	7.40E-01	0.011	3.60E-02	0.010	1.06E-01
Breast Carcinoma	TMB	Age	0.007	1.40E-03	0.008	8.14E-03	0.004	4.54E-02	0.006	2.18E-02
Cancer of Unknown Primary	TMB	Age	0.016	4.96E-06	—	—	0.015	1.69E-06	0.015	1.62E-04
Esophagogastric Carcinoma	TMB	Age	0.010	4.85E-03	0.018	1.78E-03	0.006	1.03E-01	0.006	2.52E-01
Glioma	TMB	Age	-0.002	2.37E-01	0.040	7.29E-62	-0.007	3.69E-04	-0.010	2.58E-05
Head and Neck Carcinoma	TMB	Age	0.019	2.48E-06	0.020	3.23E-05	0.018	1.03E-04	0.019	4.51E-04
Leukemia	TMB	Age	0.013	2.05E-04	—	—	—	—	—	—
Melanoma	TMB	Age	0.011	6.17E-05	0.006	4.76E-02	0.008	1.39E-02	0.009	3.95E-02
Non-Small Cell Lung Cancer	TMB	Age	0.008	3.26E-05	-0.012	2.60E-03	-0.001	4.83E-01	-0.002	3.78E-01
Prostate Cancer	TMB	Age	0.018	2.72E-03	0.038	5.07E-07	0.010	4.68E-02	0.013	1.92E-02
Renal Cell Carcinoma	TMB	Age	0.021	1.73E-07	0.034	1.48E-12	—	—	—	—
Soft Tissue Sarcoma	TMB	Age	0.017	1.59E-09	—	—	0.013	2.94E-07	0.016	2.09E-05
Pan-Cancer	TMB	Age	0.007	3.68E-25	0.018	6.20E-60	0.006	2.43E-14	0.006	9.57E-11
Bladder Cancer	TMB	Sex	-0.008	9.38E-01	-0.340	1.55E-02	-0.144	2.50E-01	-0.196	1.90E-01
Glioma	TMB	Sex	0.097	9.74E-02	0.229	6.46E-04	0.009	8.83E-01	0.020	7.86E-01
Melanoma	TMB	Sex	-0.295	1.44E-04	-0.311	1.46E-03	-0.374	2.77E-04	-0.370	4.18E-03
Renal Cell Carcinoma	TMB	Sex	0.064	4.96E-01	-0.001	9.94E-01	—	—	—	—
Pan-Cancer	TMB	Sex	-0.034	8.45E-02	-0.044	1.95E-01	-0.050	2.16E-02	-0.048	6.85E-02
Bladder Cancer	All CNB	Prostate Cancer PRS (5E-5)	-0.122	3.13E-03	0.086	1.25E-01	—	—	—	—
Colorectal Cancer	All CNB	Drinks Per Week PRS (5E-4)	-0.083	2.31E-03	-0.108	4.90E-02	—	—	—	—
Endometrial Cancer	All CNB	Education in Years PRS (5E-3)	0.131	7.43E-04	0.098	7.46E-02	—	—	—	—
Melanoma	All CNB	Smoker with Lung Cancer PRS (5E-2)	-0.160	5.65E-05	0.024	5.21E-01	—	—	—	—
Non-Hodgkin Lymphoma	All CNB	Autoimmune Disease (5E-6)	0.159	2.61E-03	—	—	—	—	—	—
Pan-Cancer	All CNB	Autoimmune Disease (5E-1)	0.026	1.41E-03	-0.008	5.82E-01	—	—	—	—
Melanoma	CNB	Education in Years PRS (5E-7)	-0.137	6.86E-04	-0.025	5.27E-01	0.008	8.74E-01	-0.035	5.35E-01
Breast Carcinoma	TMB	Renal Cell Carcinoma PRS (5E-3)	-0.080	1.82E-03	0.024	5.35E-01	0.007	8.09E-01	0.005	8.63E-01
Melanoma	TMB	Ease of Tanning PRS (5E-2)	0.168	3.18E-05	-0.003	9.56E-01	0.134	9.40E-03	0.177	4.90E-03
Non-Small Cell Lung Cancer	TMB	Cigarettes Per Day PRS (5E-1)	0.098	1.54E-06	0.093	1.07E-02	0.054	1.30E-02	0.073	3.05E-03
Non-Small Cell Lung Cancer	TMB	Smoker with Lung Cancer PRS (5E-7)	0.083	2.24E-05	0.086	1.67E-02	-0.009	6.95E-01	0.020	4.25E-01
Non-Small Cell Lung Cancer	TMB	Education in Years PRS (5E-1)	-0.065	1.35E-03	0.088	1.74E-02	-0.056	1.15E-02	-0.034	1.71E-01
Soft Tissue Sarcoma	TMB	Drinks Per Week PRS (5E-4)	0.135	1.14E-03	—	—	0.053	2.50E-01	0.003	9.65E-01
Pan-Cancer	TMB	Cigarettes Per Day PRS (5E-1)	0.028	1.06E-03	0.001	9.31E-01	0.019	4.67E-02	0.030	6.10E-03
Pan-Cancer	TMB	White Blood Cell Count PRS (5E-6)	0.025	3.37E-03	-0.027	4.72E-02	0.002	8.66E-01	0.003	8.03E-01
Pan-Cancer	TMB	Education in Years PRS (5E-0)	-0.028	1.37E-03	0.022	1.10E-01	-0.018	6.99E-02	-0.016	1.42E-01

Table 4.1: **Comparison of discoveries to previous findings.** All significant discoveries in Profile for age, sex and PRS are reported here. These associations were then tested in Tempus and TCGA using our pipeline as long as a sufficient sample size was available. We also re-analyzed the previously reported discoveries in TCGA using our pipeline in all three cohorts.

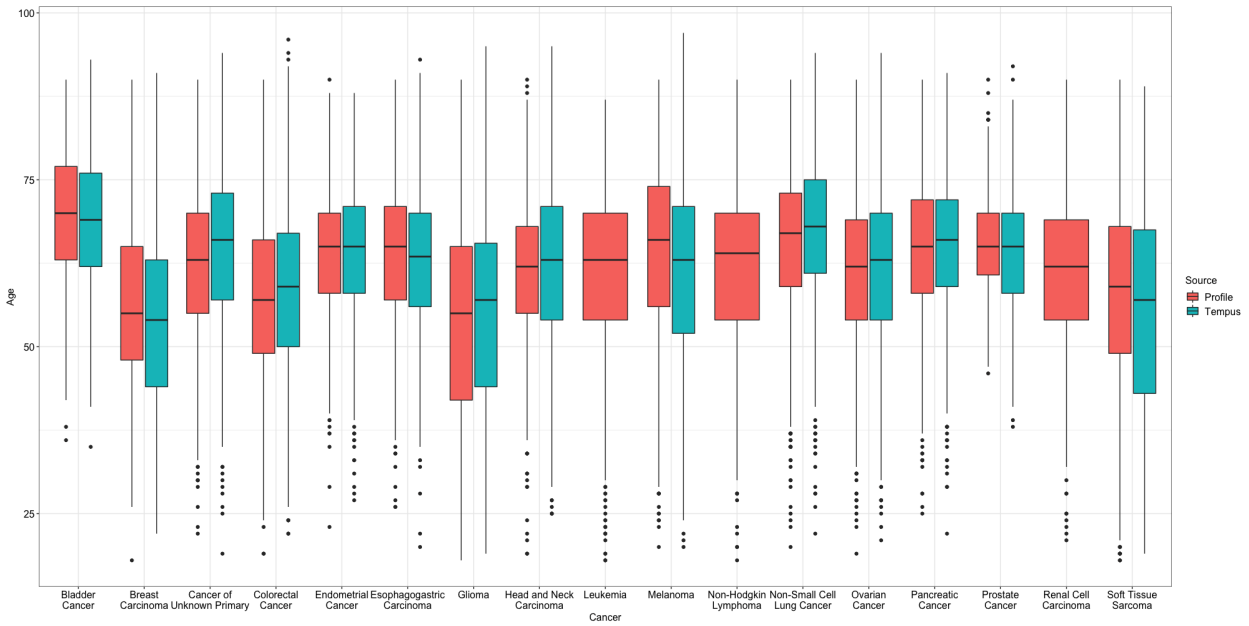


Figure 4.12: **Distribution of age across cancers.** Box-plot of the distribution of age for each cancer in Profile (red) and Tempus (blue). The box represents the interquartile range with the median value indicated within.

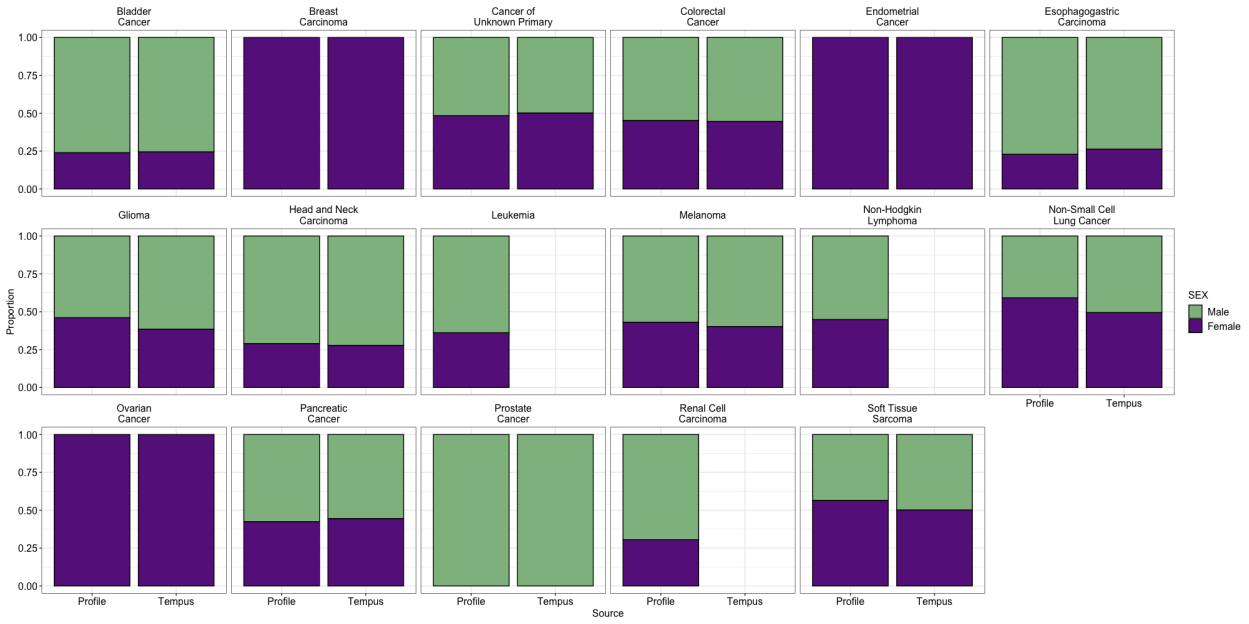


Figure 4.13: **Distribution of sex across cancers.** Bar graph of the proportion of women (purple) with each cancer in a separate panel. Within each panel Profile is on the left and Tempus on the right.

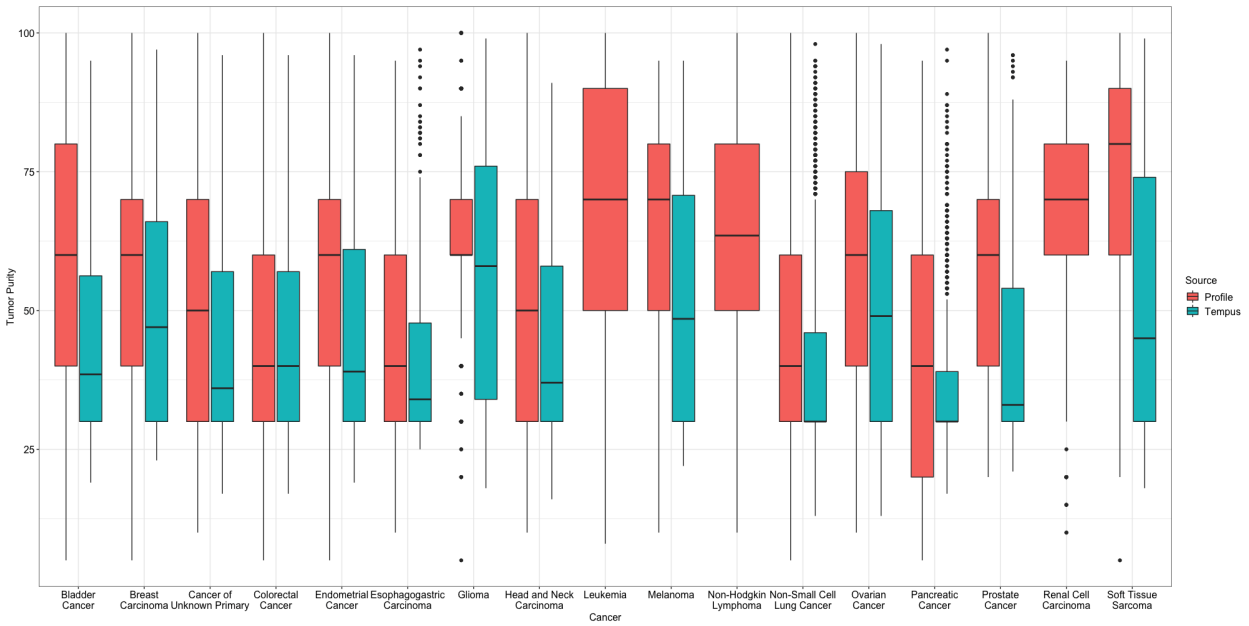


Figure 4.14: **Distribution of tumor purity across cancers.** Box-plot of the distribution of tumor purity for each cancer in Profile (red) and Tempus (blue). The box represents the interquartile range with the median value indicated within.

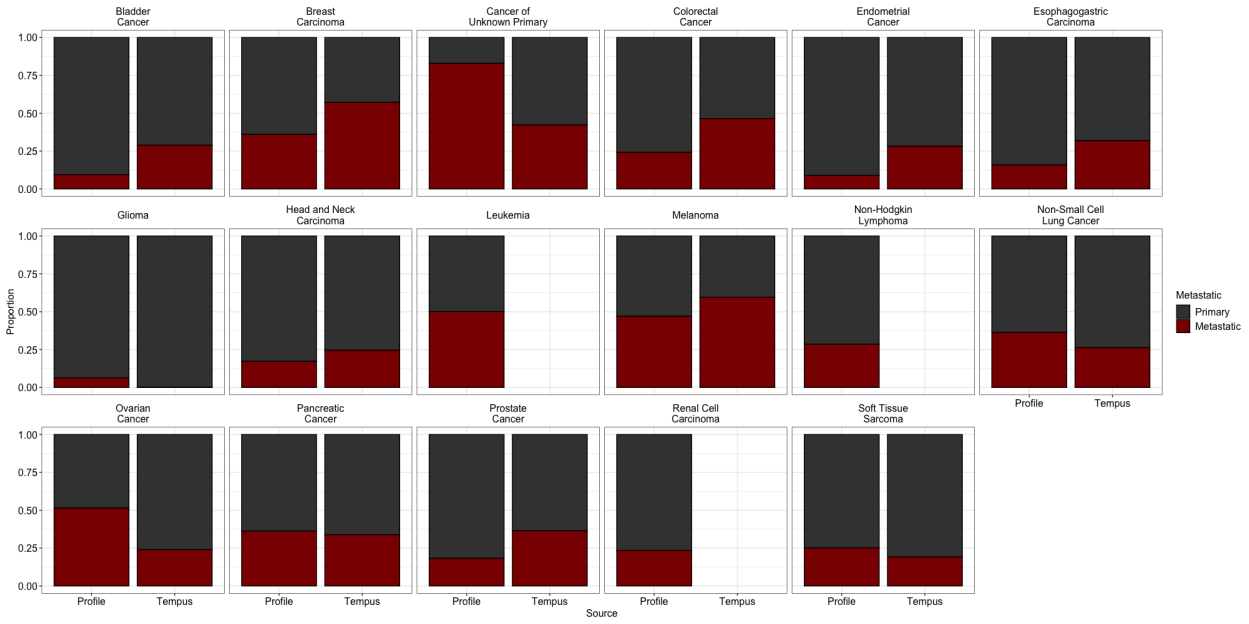


Figure 4.15: **Distribution of metastatic status across cancers.** Bar graph of the proportion of metastatic patients (red) with each cancer in a separate panel. Within each panel Profile is on the left and Tempus on the right.

CHAPTER 5

Conclusions and Future Work

With the continual growth of biobanks and genomic datasets spanning numerous phenotypes and diverse genetic ancestries, there is a pressing need for both methodological advancements as well as discoveries via meta-analyses. Here, I present two novel methods for meta-analysis as well as an exploration of determinants of somatic genetic variant burden pan-cancer. In chapter 1, I introduced a number of key concepts in the field of quantitative genetics, including single nucleotide polymorphisms (SNPs), linkage disequilibrium (LD), genome-wide association studies (GWAS), and fine-mapping. In chapter 2, I introduced a multi-trait method, PAT, that leverages the covariance structure between traits, particularly how the covariance matrix can be decomposed into genetic and environmental components. In chapter 3, I introduced a multi-ancestry method, MsCAVIAR, which utilizes the differing LD patterns between distinct ancestral backgrounds to refine the set of associated variants into a subset known as the causal set. In chapter 4, I examined a number of clinical and polygenic germline determinants of somatic variant burden both within individual cancers and pan-cancer via a meta-analysis. Through my doctoral work, I was able to contribute to our understanding of the genetic architecture of complex traits and disease by leveraging the shared information between data sources.

While my dissertation presents a number of small advancements, there are numerous future directions each of my contributions could take. One such direction based on chapter 2, would be to address the assumption that the genetic covariance structure is constant across the genome. By allowing a different covariance structure for different sets of SNPs, there could be a significant power increase as modeling local covariance structure better reflects the covariance structure between summary statistics [46, 139, 140]. Our method, however,

only considered the global estimate of genetic and environmental correlation between traits and further work is needed to quantify the impact of such modifications on both power and false discovery which others have begun to explore [84].

Another future research direction is to alter how the m-value interpretation framework presented in chapter 2 estimates the number of causal variants for the genetic covariance matrix. Currently, for association testing the genetic covariance matrix was scaled according to the polygenic model (i.e. all variants were causal). Once variants were implicated as associated, we used grid search to find the genetic covariance matrix scaling that best reflected the average effect size of independent variants. This in effect was an approximation to the number of causal variants. Further work is merited to better estimate the number of causal variants for each trait as well as the number of causal SNPs shared between traits.

For the method presented in chapter 3, MsCAVIAR currently only provides one causal set which with ρ probability contains *all* causal SNPs. This is problematic as it is not clear how many distinct causal signals are contained within the set of variants. Another method, SuSiE, provides credible sets where each set has ρ probability of containing at least one causal variant [158]. Further work is merited to extend the CAVIAR/MsCAVIAR framework, so that the signal within the causal set can be partitioned into credible sets.

Lastly, we could extend the work presented in chapter 4 by exploring further phenotypes based on somatic variants in cancer. Currently, we have only considered somatic variant burden defined genome-wide, but it is possible that gene level events or individual hotspot mutations are also genetically determined separately from somatic variant burden. Additionally, clonal and subclonal somatic variant burden may be influenced differently by clinical and germline genetics which warrants consideration.

CHAPTER 6

Bibliography

BIBLIOGRAPHY

[1] 1000 Genomes Project Consortium, A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison, H. M. Kang, J. O. Korbel, J. L. Marchini, S. McCarthy, G. A. McVean, and G. R. Abecasis, "A global reference for human genetic variation," *Nature*, vol. 526, no. 7571, pp. 68–74, Oct. 2015. 43, 47, 52, 55, 57, 81

[2] A. G. Akritas, E. K. Akritas, and G. I. Malaschonok, "Various proofs of sylvester's (determinant) identity," *Math. Comput. Simul.*, vol. 42, no. 4, pp. 585–593, Nov. 1996. 67

[3] L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, S. A. J. R. Aparicio, S. Behjati, A. V. Biankin, G. R. Bignell, N. Bolli, A. Borg, A.-L. Børresen-Dale, S. Boyault, B. Burkhardt, A. P. Butler, C. Caldas, H. R. Davies, C. Desmedt, R. Eils, J. E. Eyfjörd, J. A. Foekens, M. Greaves, F. Hosoda, B. Hutter, T. Ilicic, S. Imbeaud, M. Imielinski, N. Jäger, D. T. W. Jones, D. Jones, S. Knappskog, M. Kool, S. R. Lakhani, C. López-Otín, S. Martin, N. C. Munshi, H. Nakamura, P. A. Northcott, M. Pajic, E. Papaemmanuil, A. Paradiso, J. V. Pearson, X. S. Puente, K. Raine, M. Ramakrishna, A. L. Richardson, J. Richter, P. Rosenstiel, M. Schlesner, T. N. Schumacher, P. N. Span, J. W. Teague, Y. Totoki, A. N. J. Tutt, R. Valdés-Mas, M. M. van Buuren, L. van 't Veer, A. Vincent-Salomon, N. Waddell, L. R. Yates, Australian Pancreatic Cancer Genome Initiative, ICGC Breast Cancer Consortium, ICGC MMML-Seq Consortium, ICGC PedBrain, J. Zucman-Rossi, P. A. Futreal, U. McDermott, P. Lichter, M. Meyerson, S. M. Grimmond, R. Siebert, E. Campo, T. Shibata, S. M. Pfister, P. J. Campbell, and M. R. Stratton, "Signatures of mutational processes in human cancer," *Nature*, vol. 500, no. 7463, pp. 415–421, Aug. 2013. 76, 83

[4] F. André, M. Arnedos, A. S. Baras, J. Baselga, P. L. Bedard, M. F. Berger, M. Bierkens, F. Calvo, E. Cerami, D. Chakravarty, K. K. Dang, N. E. Davidson, D. V. Fitz, S. Dogan, R. N. DuBois, M. D. Ducar, P. Andrew, J. Gao, F. Garcia, S. Gardos, C. D. Gocke, B. E. Gross, J. Guinney, Z. J. Heins, S. Hintzen, H. Horlings, J. Hudeček, D. M. Hy-

man, S. Kamel-Reid, C. Kandoth, W. Kinyua, P. Kumari, R. Kundra, M. Ladanyi, C. Lefebvre, M. L. LeNoue-Newton, E. M. Lepisto, M. A. Levy, N. I. Lindeman, J. Lindsay, D. Liu, Z. Lu, L. E. MacConaill, I. Maurer, D. S. Maxwell, G. A. Meijer, F. Meric-Bernstam, C. M. Micheel, C. Miller, G. Mills, N. D. Moore, P. M. Nederlof, L. Omberg, J. A. Orechia, B. H. Park, T. J. Pugh, B. Reardon, B. J. Rollins, M. J. Routbort, C. L. Sawyers, D. Schrag, N. Schultz, K. R. Mills, P. Shivdasani, L. L. Siu, D. B. Solit, G. S. Sonke, J. C. Soria, P. Sripakdeevong, N. H. Stickle, T. P. Stricker, S. M. Sweeney, B. S. Taylor, J. Jelle, S. B. Thomas, V. Allen, M. Eliezer, V. T. Veer, J. Laura, V. E. Velculescu, C. Virtanen, E. E. Voest, L. L. Wang, C. Wathoo, S. Watt, C. Yu, T. V. Yu, E. Yu, A. Zehir, and H. Zhang, "AACR project GENIE: Powering precision medicine through an international consortium," *Cancer Discov.*, vol. 7, no. 8, pp. 818–831, Aug. 2017. 78

[5] D. Aran, M. Sirota, and A. J. Butte, "Systematic pan-cancer analysis of tumour purity," *Nat. Commun.*, vol. 6, no. 1, pp. 1–12, Dec. 2015. 103

[6] H. Aschard, V. Guillemot, B. Vilhjalmsson, C. J. Patel, D. Skurnik, C. J. Ye, B. Wolpin, P. Kraft, and N. Zaitlen, "Covariate selection for association screening in multiphenotype genetic studies," *Nat. Genet.*, vol. 49, no. 12, pp. 1789–1795, Dec. 2017. 8

[7] J. L. Asimit, D. B. Rainbow, M. D. Fortune, N. F. Grinberg, L. S. Wicker, and C. Wallace, "Stochastic search and joint fine-mapping increases accuracy and identifies previously unreported associations in immune-mediated diseases," *Nat. Commun.*, vol. 10, no. 1, p. 3216, Jul. 2019. 62

[8] A. Balmain, "The critical roles of somatic mutations and environmental tumor-promoting agents in cancer risk," *Nat. Genet.*, vol. 52, no. 11, pp. 1139–1143, Nov. 2020. 76

[9] R. Barroso-Sousa, E. Jain, O. Cohen, D. Kim, J. Buendia-Buendia, E. Winer, N. Lin, S. M. Tolaney, and N. Wagle, "Prevalence and mutational determinants of high tumor

mutation burden in breast cancer," *Ann. Oncol.*, vol. 31, no. 3, pp. 387–394, Mar. 2020. 77

[10] N. Beaubier, M. Bontrager, R. Huether, C. Igartua, D. Lau, R. Tell, A. M. Bobe, S. Bush, A. L. Chang, D. C. Hoskinson, A. A. Khan, E. Kudalkar, B. D. Leibowitz, A. Lozachmeur, J. Michuda, J. Parsons, J. F. Perera, A. Salahudeen, K. P. Shah, T. Taxter, W. Zhu, and K. P. White, "Integrated genomic profiling expands clinical options for patients with cancer," *Nat. Biotechnol.*, vol. 37, no. 11, pp. 1351–1360, Sep. 2019. 3, 79, 99, 102

[11] A. H. Beecham, N. A. Patsopoulos, D. K. Xifara, M. F. Davis, A. Kemppinen, C. Cotsapas, T. S. Shah, C. Spencer, D. Booth, A. Goris *et al.*, "Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis," *Nat Genet.*, vol. 45, no. 11, p. 1353, 2013. 45

[12] S. Benafif, Z. Kote-Jarai, R. A. Eeles, and PRACTICAL Consortium, "A review of prostate cancer Genome-Wide association studies (GWAS)," *Cancer Epidemiol. Biomarkers Prev.*, vol. 27, no. 8, pp. 845–857, Aug. 2018. 76

[13] C. Benner, C. C. A. Spencer, A. S. Havulinna, V. Salomaa, S. Ripatti, and M. Pirinen, "FINEMAP: efficient variable selection using summary data from genome-wide association studies," *Bioinformatics*, vol. 32, no. 10, pp. 1493–1501, May 2016. 4, 46, 52, 62, 66

[14] S. Bhattacharjee, P. Rajaraman, K. B. Jacobs, W. A. Wheeler, B. S. Melin, P. Hartge, GliomaScan Consortium, M. Yeager, C. C. Chung, S. J. Chanock, and N. Chatterjee, "A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits," *Am. J. Hum. Genet.*, vol. 90, no. 5, pp. 821–835, May 2012. 6, 7, 15

[15] S. Bolormaa, J. E. Pryce, A. Reverter, Y. Zhang, W. Barendse, K. Kemper, B. Tier, K. Savin, B. J. Hayes, and M. E. Goddard, "A multi-trait, meta-analysis for detecting

pleiotropic polymorphisms for stature, fatness and reproduction in beef cattle," *PLoS Genet.*, vol. 10, no. 3, p. e1004198, Mar. 2014. 6

[16] A. R. Buckley, K. A. Standish, K. Bhutani, T. Ideker, R. S. Lasken, H. Carter, O. Harismendy, and N. J. Schork, "Pan-cancer analysis reveals technical artifacts in TCGA germline variant calls," *BMC Genomics*, vol. 18, no. 1, p. 458, Jun. 2017. 77, 100

[17] B. Bulik-Sullivan, H. K. Finucane, V. Anttila, A. Gusev, F. R. Day, P.-R. Loh, ReproGen Consortium, Psychiatric Genomics Consortium, Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium 3, L. Duncan, J. R. B. Perry, N. Patterson, E. B. Robinson, M. J. Daly, A. L. Price, and B. M. Neale, "An atlas of genetic correlations across human diseases and traits," *Nat. Genet.*, vol. 47, no. 11, pp. 1236–1241, Nov. 2015. 24, 43

[18] B. K. Bulik-Sullivan, P.-R. Loh, H. K. Finucane, S. Ripke, J. Yang, Schizophrenia Working Group of the Psychiatric Genomics Consortium, N. Patterson, M. J. Daly, A. L. Price, and B. M. Neale, "LD score regression distinguishes confounding from polygenicity in genome-wide association studies," *Nat. Genet.*, vol. 47, no. 3, pp. 291–295, Mar. 2015. 24, 43

[19] R. Büttner, J. W. Longshore, F. López-Ríos, S. Merkelbach-Bruse, N. Normanno, E. Rouleau, and F. Penault-Llorca, "Implementing TMB measurement in clinical practice: considerations on assay requirements," *ESMO Open*, vol. 4, no. 1, p. e000442, Jan. 2019. 76

[20] J. Carrot-Zhang, N. Chambwe, J. S. Damrauer, T. A. Knijnenburg, A. G. Robertson, C. Yau, W. Zhou, A. C. Berger, K.-L. Huang, J. Y. Newberg, R. J. Mashl, A. Romanel, R. W. Sayaman, F. Demichelis, I. Felau, G. M. Frampton, S. Han, K. A. Hoadley, A. Kemal, P. W. Laird, A. J. Lazar, X. Le, N. Oak, H. Shen, C. K. Wong, J. C. Zenklusen, E. Ziv, Cancer Genome Atlas Analysis Network, A. D. Cherniack, and R. Beroukhim, "Comprehensive analysis of genetic ancestry and its molecular correlates in cancer," *Cancer Cell*, vol. 37, no. 5, pp. 639–654.e6, May 2020. 99

[21] Z. R. Chalmers, C. F. Connelly, D. Fabrizio, L. Gay, S. M. Ali, R. Ennis, A. Schrock, B. Campbell, A. Shlien, J. Chmielecki, F. Huang, Y. He, J. Sun, U. Tabori, M. Kennedy, D. S. Lieber, S. Roels, J. White, G. A. Otto, J. S. Ross, L. Garraway, V. A. Miller, P. J. Stephens, and G. M. Frampton, "Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden," *Genome Med.*, vol. 9, no. 1, p. 34, Apr. 2017. 76, 83

[22] C.-Y. Chen, S. Pollack, D. J. Hunter, J. N. Hirschhorn, P. Kraft, and A. L. Price, "Improved ancestry inference using weights from external reference panels," *Bioinformatics*, vol. 29, no. 11, pp. 1399–1406, Jun. 2013. 101, 104

[23] S. Chen, E. S. Iversen, T. Friebel, D. Finkelstein, B. L. Weber, A. Eisen, L. E. Peterson, J. M. Schildkraut, C. Isaacs, B. N. Peshkin, C. Corio, L. Leondaridis, G. Tomlinson, D. Dutson, R. Kerber, C. I. Amos, L. C. Strong, D. A. Berry, D. M. Euhus, and G. Parmigiani, "Characterization of BRCA1 and BRCA2 mutations in a large united states sample," *J. Clin. Oncol.*, vol. 24, no. 6, pp. 863–871, Feb. 2006. 76

[24] S. Chen, W. Wang, S. Lee, K. Nafa, J. Lee, K. Romans, P. Watson, S. B. Gruber, D. Euhus, K. W. Kinzler, J. Jass, S. Gallinger, N. M. Lindor, G. Casey, N. Ellis, F. M. Giardiello, K. Offit, G. Parmigiani, and Colon Cancer Family Registry, "Prediction of germline mutations and cancer risk in the lynch syndrome," *JAMA*, vol. 296, no. 12, pp. 1479–1487, Sep. 2006. 4, 76

[25] W. Chen, B. R. Larrabee, I. G. Ovsyannikova, R. B. Kennedy, I. H. Haralambieva, G. A. Poland, and D. J. Schaid, "Fine mapping causal variants with an approximate bayesian method using marginal test statistics," *Genetics*, vol. 200, no. 3, pp. 719–736, Jul. 2015. 46, 66

[26] D. T. Cheng, T. N. Mitchell, A. Zehir, R. H. Shah, R. Benayed, A. Syed, R. Chandramohan, Z. Y. Liu, H. H. Won, S. N. Scott, A. R. Brannon, C. O'Reilly, J. Sadowska, J. Casanova, A. Yannes, J. F. Hechtman, J. Yao, W. Song, D. S. Ross, A. Oultache, S. Dogan, L. Borsu, M. Hameed, K. Nafa, M. E. Arcila, M. Ladanyi, and M. F. Berger,

"Memorial sloan Kettering-Integrated mutation profiling of actionable cancer targets (MSK-IMPACT): A hybridization Capture-Based Next-Generation sequencing clinical assay for solid tumor molecular oncology," *J. Mol. Diagn.*, vol. 17, no. 3, pp. 251–264, May 2015. 78

[27] K. Chesmore, J. Bartlett, and S. M. Williams, "The ubiquity of pleiotropy in human disease," *Hum. Genet.*, vol. 137, no. 1, pp. 39–44, Jan. 2018. 4, 6

[28] J.-H. Choi, S.-E. Hong, and H. G. Woo, "Pan-cancer analysis of systematic batch effects on somatic sequence variations," *BMC Bioinformatics*, vol. 18, no. 1, p. 211, Apr. 2017. 77, 100

[29] M. B. Cook, S. M. Dawsey, N. D. Freedman, P. D. Inskip, S. M. Wichner, S. M. Quraishi, S. S. Devesa, and K. A. McGlynn, "Sex disparities in cancer incidence by period and age," *Cancer Epidemiol. Biomarkers Prev.*, vol. 18, no. 4, pp. 1174–1182, Apr. 2009. 77

[30] M. B. Cook, K. A. McGlynn, S. S. Devesa, N. D. Freedman, and W. F. Anderson, "Sex disparities in cancer mortality and survival," *Cancer Epidemiol. Biomarkers Prev.*, vol. 20, no. 8, pp. 1629–1637, Aug. 2011. 77, 92, 94

[31] P. J. Cook, R. Doll, and S. A. Fellingham, "A mathematical model for the age distribution of cancer in man," *Int. J. Cancer*, vol. 4, no. 1, pp. 93–112, Jan. 1969. 77

[32] Cross-Disorder Group of the Psychiatric Genomics Consortium, "Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis," *Lancet*, vol. 381, no. 9875, pp. 1371–1379, Apr. 2013. 6

[33] R. W. Davies, J. Flint, S. Myers, and R. Mott, "Rapid genotype imputation from sequence without reference panels," *Nat. Genet.*, vol. 48, no. 8, pp. 965–969, Jul. 2016. 101

[34] T. Davoli, H. Uno, E. C. Wooten, and S. J. Elledge, "Tumor aneuploidy correlates with

markers of immune evasion and with reduced response to immunotherapy," *Science*, vol. 355, no. 6322, Jan. 2017. 77, 92, 95

[35] J. P. de Magalhães, "How ageing processes influence cancer," *Nat. Rev. Cancer*, vol. 13, no. 5, pp. 357–365, May 2013. 77

[36] DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, Asian Genetic Epidemiology Network Type 2 Diabetes (AGEN-T2D) Consortium, South Asian Type 2 Diabetes (SAT2D) Consortium, Mexican American Type 2 Diabetes (MAT2D) Consortium, Type 2 Diabetes Genetic Exploration by Nex-generation sequencing in muylti-Ethnic Samples (T2D-GENES) Consortium, A. Mahajan, M. J. Go, W. Zhang, J. E. Below, K. J. Gaulton, T. Ferreira, M. Horikoshi, A. D. Johnson, M. C. Y. Ng, I. Prokopenko, D. Saleheen, X. Wang, E. Zeggini, G. R. Abecasis, L. S. Adair, P. Almgren, M. Atalay, T. Aung, D. Baldassarre, B. Balkau, Y. Bao, A. H. Barnett, I. Barroso, A. Basit, L. F. Been, J. Beilby, G. I. Bell, R. Benediktsson, R. N. Bergman, B. O. Boehm, E. Boerwinkle, L. L. Bonnycastle, N. Burtt, Q. Cai, H. Campbell, J. Carey, S. Cauchi, M. Caulfield, J. C. N. Chan, L.-C. Chang, T.-J. Chang, Y.-C. Chang, G. Charpentier, C.-H. Chen, H. Chen, Y.-T. Chen, K.-S. Chia, M. Chidambaram, P. S. Chines, N. H. Cho, Y. M. Cho, L.-M. Chuang, F. S. Collins, M. C. Cornelis, D. J. Couper, A. T. Crenshaw, R. M. van Dam, J. Danesh, D. Das, U. de Faire, G. Dedoussis, P. Deloukas, A. S. Dimas, C. Dina, A. S. Doney, P. J. Donnelly, M. Dorkhan, C. van Duijn, J. Dupuis, S. Edkins, P. Elliott, V. Emilsson, R. Erbel, J. G. Eriksson, J. Escobedo, T. Esko, E. Eury, J. C. Florez, P. Fontanillas, N. G. Forouhi, T. Forsen, C. Fox, R. M. Fraser, T. M. Frayling, P. Froguel, P. Frossard, Y. Gao, K. Gertow, C. Gieger, B. Gigante, H. Grallert, G. B. Grant, L. C. Grrop, C. J. Groves, E. Grundberg, C. Guiducci, A. Hamsten, B.-G. Han, K. Hara, N. Hassanali, A. T. Hattersley, C. Hayward, A. K. Hedman, C. Herder, A. Hofman, O. L. Holmen, K. Hovingh, A. B. Hreidarsson, C. Hu, F. B. Hu, J. Hui, S. E. Humphries, S. E. Hunt, D. J. Hunter, K. Hveem, Z. I. Hydrie, H. Ikegami, T. Illig, E. Ingelsson, M. Islam, B. Isomaa, A. U. Jackson, T. Jafar, A. James, W. Jia, K.-H. Jöckel, A. Jonsson,

J. B. M. Jowett, T. Kadowaki, H. M. Kang, S. Kanoni, W. H. L. Kao, S. Kathiresan, N. Kato, P. Katulanda, K. M. Keinanen-Kiukaanniemi, A. M. Kelly, H. Khan, K.-T. Khaw, C.-C. Khor, H.-L. Kim, S. Kim, Y. J. Kim, L. Kinnunen, N. Klopp, A. Kong, E. Korpi-Hyövälti, S. Kowlessur, P. Kraft, J. Kravic, M. M. Kristensen, S. Krithika, A. Kumar, J. Kumate, J. Kuusisto, S. H. Kwak, M. Laakso, V. Lagou, T. A. Lakka, C. Langenberg, C. Langford, R. Lawrence, K. Leander, J.-M. Lee, N. R. Lee, M. Li, X. Li, Y. Li, J. Liang, S. Liju, W.-Y. Lim, L. Lind, C. M. Lindgren, E. Lindholm, C.-T. Liu, J. J. Liu, S. Lobbens, J. Long, R. J. F. Loos, W. Lu, J. Luan, V. Lyssenko, R. C. W. Ma, S. Maeda, R. Mägi, S. Männisto, D. R. Matthews, J. B. Meigs, O. Melander, A. Metspalu, J. Meyer, G. Mirza, E. Mihailov, S. Moebus, V. Mohan, K. L. Mohlke, A. D. Morris, T. W. Mühleisen, M. Müller-Nurasyid, B. Musk, J. Nakamura, E. Nakashima, P. Navarro, P.-K. Ng, A. C. Nica, P. M. Nilsson, I. Njølstad, M. M. Nöthen, K. Ohnaka, T. H. Ong, K. R. Owen, C. N. A. Palmer, J. S. Pankow, K. S. Park, M. Parkin, S. Pechlivanis, N. L. Pedersen, L. Peltonen, J. R. B. Perry, A. Peters, J. M. Pinidiyapathirage, C. G. Platou, S. Potter, J. F. Price, L. Qi, V. Radha, L. Rallidis, A. Rasheed, W. Rathman, R. Rauramaa, S. Raychaudhuri, N. W. Rayner, S. D. Rees, E. Rehnberg, S. Ripatti, N. Robertson, M. Roden, E. J. Rossin, I. Rudan, D. Rybin, T. E. Saaristo, V. Salomaa, J. Saltevo, M. Samuel, D. K. Sanghera, J. Saramies, J. Scott, L. J. Scott, R. A. Scott, A. V. Segrè, J. Sehmi, B. Sennblad, N. Shah, S. Shah, A. S. Shera, X. O. Shu, A. R. Shuldiner, G. Sigursson, E. Sijbrands, A. Silveira, X. Sim, S. Sivapalaratnam, K. S. Small, W. Y. So, A. Stančáková, K. Stefansson, G. Steinbach, V. Steinthorsdottir, K. Stirrups, R. J. Strawbridge, H. M. Stringham, Q. Sun, C. Suo, A.-C. Syvänen, R. Takayanagi, F. Takeuchi, W. T. Tay, T. M. Teslovich, B. Thorand, G. Thorleifsson, U. Thorsteinsdottir, E. Tikkanen, J. Trakalo, E. Tremoli, M. D. Trip, F. J. Tsai, T. Tuomi, J. Tuomilehto, A. G. Uitterlinden, A. Valladares-Salgado, S. Vedantam, F. Veglia, B. F. Voight, C. Wang, N. J. Wareham, R. Wennauer, A. R. Wickremasinghe, T. Wilsgaard, J. F. Wilson, S. Wiltshire, W. Winckler, T. Y. Wong, A. R. Wood, J.-Y. Wu, Y. Wu, K. Yamamoto, T. Yamauchi, M. Yang, L. Yengo, M. Yokota, R. Young, D. Zabaneh, F. Zhang, R. Zhang, W. Zheng, P. Z. Zimmet,

D. Altshuler, D. W. Bowden, Y. S. Cho, N. J. Cox, M. Cruz, C. L. Hanis, J. Kooner, J.-Y. Lee, M. Seielstad, Y. Y. Teo, M. Boehnke, E. J. Parra, J. C. Chambers, E. S. Tai, M. I. McCarthy, and A. P. Morris, "Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility," *Nat. Genet.*, vol. 46, no. 3, pp. 234–244, Mar. 2014. 46

[37] L. Ding, M. H. Bailey, E. Porta-Pardo, V. Thorsson, A. Colaprico, D. Bertrand, D. L. Gibbs, A. Weerasinghe, K.-L. Huang, C. Tokheim, I. Cortés-Ciriano, R. Jayasinghe, F. Chen, L. Yu, S. Sun, C. Olsen, J. Kim, A. M. Taylor, A. D. Cherniack, R. Akbani, C. Suphavilai, N. Nagarajan, J. M. Stuart, G. B. Mills, M. A. Wyczalkowski, B. G. Vincent, C. M. Hutter, J. C. Zenklusen, K. A. Hoadley, M. C. Wendl, L. Shmulevich, A. J. Lazar, D. A. Wheeler, G. Getz, and Cancer Genome Atlas Research Network, "Perspective on oncogenic processes at the end of the beginning of cancer genomics," *Cell*, vol. 173, no. 2, pp. 305–320.e10, Apr. 2018. 103

[38] G. S. Dite, M. A. Jenkins, M. C. Southey, J. S. Hocking, G. G. Giles, M. R. E. McCredie, D. J. Venter, and J. L. Hopper, "Familial risks, early-onset breast cancer, and BRCA1 and BRCA2 germline mutations," *J. Natl. Cancer Inst.*, vol. 95, no. 6, pp. 448–457, Mar. 2003. 76

[39] G. W. Dorn and S. Cresci, "Genome-wide association studies of coronary artery disease and heart failure: where are we going?" *Pharmacogenomics*, vol. 10, no. 2, pp. 213–223, feb 2009. 6

[40] M. G. Dunlop, S. M. Farrington, A. D. Carothers, A. H. Wyllie, L. Sharp, J. Burn, B. Liu, K. W. Kinzler, and B. Vogelstein, "Cancer risk associated with germline DNA mismatch repair gene mutations," *Hum. Mol. Genet.*, vol. 6, no. 1, pp. 105–110, Jan. 1997. 4, 76

[41] G. Edgren, L. Liang, H.-O. Adami, and E. T. Chang, "Enigmatic sex disparities in cancer incidence," *Eur. J. Epidemiol.*, vol. 27, no. 3, pp. 187–196, Mar. 2012. 77

[42] E. Eskin, "Discovering genes involved in disease and the mystery of missing heritability," *Commun. ACM*, vol. 58, no. 10, pp. 80–87, Sep. 2015. 6, 35

[43] M. Fallah, X. Liu, J. Ji, A. Försti, K. Sundquist, and K. Hemminki, "Autoimmune diseases associated with non-hodgkin lymphoma: a nationwide cohort study," *Ann. Oncol.*, vol. 25, no. 10, pp. 2025–2030, Oct. 2014. 91

[44] L. Fancello, S. Gandini, P. G. Pelicci, and L. Mazzarella, "Tumor mutational burden quantification from targeted gene panels: major advancements and challenges," *J Immunother Cancer*, vol. 7, no. 1, p. 183, Jul. 2019. 76

[45] L. L. Faye, M. J. Machiela, P. Kraft, S. B. Bull, and L. Sun, "Re-ranking sequencing variants in the post-GWAS era for accurate causal variant identification," *PLoS Genet.*, vol. 9, no. 8, p. e1003609, Aug. 2013. 45

[46] H. K. Finucane, B. Bulik-Sullivan, A. Gusev, G. Trynka, Y. Reshef, P.-R. Loh, V. Anttila, H. Xu, C. Zang, K. Farh, S. Ripke, F. R. Day, ReproGen Consortium, Schizophrenia Working Group of the Psychiatric Genomics Consortium, RACI Consortium, S. Purcell, E. Stahl, S. Lindstrom, J. R. B. Perry, Y. Okada, S. Raychaudhuri, M. J. Daly, N. Patterson, B. M. Neale, and A. L. Price, "Partitioning heritability by functional annotation using genome-wide association summary statistics," *Nat. Genet.*, vol. 47, no. 11, pp. 1228–1235, Nov. 2015. 34, 111

[47] S. A. Forbes, D. Beare, H. Boutselakis, S. Bamford, N. Bindal, J. Tate, C. G. Cole, S. Ward, E. Dawson, L. Ponting, R. Stefancsik, B. Harsha, C. Y. Kok, M. Jia, H. Jubb, Z. Sondka, S. Thompson, T. De, and P. J. Campbell, "COSMIC: somatic cancer genetics at high-resolution," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D777–D783, Jan. 2017. 4, 76

[48] L. G. Fritsche, W. Igl, J. N. C. Bailey, F. Grassmann, S. Sengupta, J. L. Bragg-Gresham, K. P. Burdon, S. J. Hebbring, C. Wen, M. Gorski, I. K. Kim, D. Cho, D. Zack, E. Souied, H. P. N. Scholl, E. Bala, K. E. Lee, D. J. Hunter, R. J. Sardell,

P. Mitchell, J. E. Merriam, V. Cipriani, J. D. Hoffman, T. Schick, Y. T. E. Lechanteur, R. H. Guymer, M. P. Johnson, Y. Jiang, C. M. Stanton, G. H. S. Buitendijk, X. Zhan, A. M. Kwong, A. Boleda, M. Brooks, L. Gieser, R. Ratnapriya, K. E. Branham, J. R. Foerster, J. R. Heckenlively, M. I. Othman, B. J. Vote, H. H. Liang, E. Souzeau, I. L. McAllister, T. Isaacs, J. Hall, S. Lake, D. A. Mackey, I. J. Constable, J. E. Craig, T. E. Kitchner, Z. Yang, Z. Su, H. Luo, D. Chen, H. Ouyang, K. Flagg, D. Lin, G. Mao, H. Ferreyra, K. Stark, C. N. von Strachwitz, A. Wolf, C. Brandl, G. Rudolph, M. Olden, M. A. Morrison, D. J. Morgan, M. Schu, J. Ahn, G. Silvestri, E. E. Tsironi, K. H. Park, L. A. Farrer, A. Orlin, A. Brucker, M. Li, C. A. Curcio, S. Mohand-Saïd, J.-A. Sahel, I. Audo, M. Benchaboune, A. J. Cree, C. A. Rennie, S. V. Goverdhan, M. Grunin, S. Hagbi-Levi, P. Campochiaro, N. Katsanis, F. G. Holz, F. Blond, H. Blanché, J.-F. Deleuze, R. P. Igo, Jr, B. Truitt, N. S. Peachey, S. M. Meuer, C. E. Myers, E. L. Moore, R. Klein, M. A. Hauser, E. A. Postel, M. D. Courtenay, S. G. Schwartz, J. L. Kovach, W. K. Scott, G. Liew, A. G. Tan, B. Gopinath, J. C. Merriam, R. T. Smith, J. C. Khan, H. Shahid, A. T. Moore, J. A. McGrath, R. Laux, M. A. Brantley, Jr, A. Agarwal, L. Ersoy, A. Caramoy, T. Langmann, N. T. M. Saksens, E. K. de Jong, C. B. Hoyng, M. S. Cain, A. J. Richardson, T. M. Martin, J. Blangero, D. E. Weeks, B. Dhillon, C. M. van Duijn, K. F. Doheny, J. Romm, C. C. W. Klaver, C. Hayward, M. B. Gorin, M. L. Klein, P. N. Baird, A. I. den Hollander, S. Fauser, J. R. W. Yates, R. Allikmets, J. J. Wang, D. A. Schaumberg, B. E. K. Klein, S. A. Hagstrom, I. Chowers, A. J. Lotery, T. Léveillard, K. Zhang, M. H. Brilliant, A. W. Hewitt, A. Swaroop, E. Y. Chew, M. A. Pericak-Vance, M. DeAngelis, D. Stambolian, J. L. Haines, S. K. Iyengar, B. H. F. Weber, G. R. Abecasis, and I. M. Heid, "A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants," *Nat. Genet.*, vol. 48, no. 2, pp. 134–143, Feb. 2016. 45

[49] N. A. Furlotte and E. Eskin, "Efficient multiple-trait association and estimation of genetic correlation using the matrix-variate linear mixed model," *Genetics*, vol. 200, no. 1, pp. 59–68, May 2015. 6

[50] L. Gai and E. Eskin, "Finding associated variants in genome-wide association studies on multiple traits," *Bioinformatics*, vol. 34, no. 13, pp. i467–i474, Jul. 2018. 6

[51] H. Gao, Y. Wu, T. Zhang, Y. Wu, L. Jiang, J. Zhan, J. Li, and R. Yang, "Multiple-trait genome-wide association study based on principal component analysis for residual covariance matrix," *Heredity*, vol. 114, no. 4, p. 428, Apr. 2015. 6

[52] E. P. Garcia, A. Minkovsky, Y. Jia, M. D. Ducar, P. Shivdasani, X. Gong, A. H. Ligon, L. M. Sholl, F. C. Kuo, L. E. MacConaill, N. I. Lindeman, and F. Dong, "Validation of OncoPanel: A targeted Next-Generation sequencing assay for the detection of somatic variants in cancer," *Arch. Pathol. Lab. Med.*, vol. 141, no. 6, pp. 751–758, Jun. 2017. 3, 78, 101

[53] E. Giat, M. Ehrenfeld, and Y. Shoenfeld, "Cancer and autoimmune diseases," *Autoimmun. Rev.*, vol. 16, no. 10, pp. 1049–1057, Oct. 2017. 91

[54] D. E. Goldgar, D. F. Easton, L. A. Cannon-Albright, and M. H. Skolnick, "Systematic population-based assessment of cancer risk in first-degree relatives of cancer probands," *J. Natl. Cancer Inst.*, vol. 86, no. 21, pp. 1600–1608, Nov. 1994. 76

[55] A. Goodman, S. Kato, L. Bazhenova, S. Patel, G. Frampton, V. Miller, P. Stephens, G. Daniels, and R. Kurzrock, "Tumor mutational burden as an independent predictor of response to immunotherapy in diverse cancers," *Mol. Cancer Ther.*, 2017. 76

[56] J. Gratten and P. M. Visscher, "Genetic pleiotropy in complex traits and diseases: implications for genomic medicine," *Genome Med.*, vol. 8, no. 1, p. 78, Jul. 2016. 4, 6

[57] P. N. Gray, C. L. M. Dunlop, and A. M. Elliott, "Not all next generation sequencing diagnostics are created equal: Understanding the nuances of solid tumor assay design for somatic mutation detection," *Cancers*, vol. 7, no. 3, pp. 1313–1332, Jul. 2015. 76

[58] C. Greenman, P. Stephens, R. Smith, G. L. Dalgliesh, C. Hunter, G. Bignell, H. Davies, J. Tague, A. Butler, C. Stevens, S. Edkins, S. O'Meara, I. Vastrik, E. Schmidt, T. Avis,

S. Barthorpe, G. Bhamra, G. Buck, B. Choudhury, J. Clements, J. Cole, E. Dicks, S. Forbes, K. Gray, K. Halliday, R. Harrison, K. Hills, J. Hinton, A. Jenkinson, D. Jones, A. Menzies, T. Mironenko, J. Perry, K. Raine, D. Richardson, R. Shepherd, A. Small, C. Tofts, J. Varian, T. Webb, S. West, S. Widaa, A. Yates, D. P. Cahill, D. N. Louis, P. Goldstraw, A. G. Nicholson, F. Brasseur, L. Looijenga, B. L. Weber, Y. Chiew, A. deFazio, M. F. Greaves, A. R. Green, P. Campbell, E. Birney, D. F. Easton, G. Chenevix-Trench, M. Tan, S. K. Khoo, B. T. Teh, S. T. Yuen, S. Y. Leung, R. Wooster, F. P. A, and M. R. Stratton, "Patterns of somatic mutation in human cancer genomes," *Nature*, p. 153, 2007. 4, 76

[59] S. Gupta, M. Artomov, W. Goggins, M. Daly, and H. Tsao, "Gender disparity and mutation burden in metastatic melanoma," *J. Natl. Cancer Inst.*, vol. 107, no. 11, Nov. 2015. 83

[60] A. Gusev, S. Groha, K. Taraszka, Y. R. Semenov, and N. Zaitlen, "Constructing germline research cohorts from the discarded reads of clinical tumor sequences," *Genome Med.*, vol. 13, no. 1, p. 179, Nov. 2021. 3, 4, 99, 101

[61] M. J. Hall, A. D. Forman, R. Pilarski, G. Wiesner, and V. N. Giri, "Gene panel testing for inherited cancer risk," *J. Natl. Compr. Canc. Netw.*, vol. 12, no. 9, pp. 1339–1346, Sep. 2014. 76

[62] B. Han and E. Eskin, "Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies," *Am. J. Hum. Genet.*, vol. 88, no. 5, pp. 586–598, May 2011. 46

[63] ——, "Interpreting meta-analyses of genome-wide association studies," *PLoS Genet.*, vol. 8, no. 3, p. e1002555, Mar. 2012. 7, 9, 14, 40

[64] C. Harding, F. Pompei, and R. Wilson, "Peak and decline in cancer incidence, mortality, and prevalence at old ages," *Cancer*, vol. 118, no. 5, pp. 1371–1386, Mar. 2012. 77

[65] H. V. Henderson and S. R. Searle, "On deriving the inverse of a sum of matrices," *SIAM Rev. Soc. Ind. Appl. Math.*, vol. 23, no. 1, pp. 53–60, Jan. 1981. 66

[66] H. Hieronymus, R. Murali, A. Tin, K. Yadav, W. Abida, H. Moller, D. Berney, H. Scher, B. Carver, P. Scardino, N. Schultz, B. Taylor, A. Vickers, J. Cuzick, and C. L. Sawyers, "Tumor copy number alteration burden is a pan-cancer prognostic factor associated with recurrence and death," *Elife*, vol. 7, Sep. 2018. 77, 92

[67] J. H. J. Hoeijmakers, "DNA damage, aging, and cancer," *N. Engl. J. Med.*, vol. 361, no. 15, pp. 1475–1485, Oct. 2009. 77

[68] F. Hormozdiari, E. Kostem, E. Y. Kang, B. Pasaniuc, and E. Eskin, "Identifying causal variants at loci with multiple signals of association," *Genetics*, vol. 198, no. 2, pp. 497–508, Oct. 2014. 4, 45, 46, 50, 52, 62, 71

[69] M. Ikeda, A. Takahashi, Y. Kamatani, Y. Okahisa, H. Kunugi, N. Mori, T. Sasaki, T. Ohmori, Y. Okamoto, H. Kawasaki, S. Shimodera, T. Kato, H. Yoneda, R. Yoshimura, M. Iyo, K. Matsuda, M. Akiyama, K. Ashikawa, K. Kashiwase, K. Tokunaga, K. Kondo, T. Saito, A. Shimasaki, K. Kawase, T. Kitajima, K. Matsuo, M. Itokawa, T. Someya, T. Inada, R. Hashimoto, T. Inoue, K. Akiyama, H. Tanii, H. Arai, S. Kanba, N. Ozaki, I. Kusumi, T. Yoshikawa, M. Kubo, and N. Iwata, "A genome-wide association study identifies two novel susceptibility loci and trans population polygenicity associated with bipolar disorder," *Mol. Psychiatry*, vol. 23, no. 3, pp. 639–647, Mar. 2018. 45

[70] International HapMap Consortium, "The international HapMap project," *Nature*, vol. 426, no. 6968, pp. 789–796, Dec. 2003. 29, 35, 47

[71] L. Jiang, Z. Zheng, T. Qi, K. E. Kemper, N. R. Wray, P. M. Visscher, and J. Yang, "A resource-efficient tool for mixed model association analysis of large-scale data," *Nat. Genet.*, vol. 51, no. 12, pp. 1749–1755, Nov. 2019. 50

[72] S. Jones, V. Anagnostou, K. Lytle, S. Parpart-Li, M. Nesselbush, D. R. Riley, M. Shukla, B. Chesnick, M. Kadan, E. Papp, K. G. Galens, D. Murphy, T. Zhang, L. Kann, M. Sausen, S. V. Angiuoli, L. A. Diaz, Jr, and V. E. Velculescu, "Personalized genomic analyses for cancer mutation discovery and interpretation," *Sci. Transl. Med.*, vol. 7, no. 283, p. 283ra53, Apr. 2015. 79

[73] M. Kanai, M. Akiyama, A. Takahashi, N. Matoba, Y. Momozawa, M. Ikeda, N. Iwata, S. Ikegawa, M. Hirata, K. Matsuda, M. Kubo, Y. Okada, and Y. Kamatani, "Genetic analysis of quantitative traits in the japanese population links cell types to complex human diseases," *Nat. Genet.*, vol. 50, no. 3, pp. 390–400, Mar. 2018. 51

[74] D. Karasik and D. P. Kiel, "Evidence for pleiotropic factors in genetics of the musculoskeletal system," *Bone*, vol. 46, no. 5, pp. 1226–1237, May 2010. 6

[75] G. Kichaev and B. Pasaniuc, "Leveraging Functional-Annotation data in trans-ethnic Fine-Mapping studies," *Am. J. Hum. Genet.*, vol. 97, no. 2, pp. 260–271, Aug. 2015. 46, 50

[76] G. Kichaev, W.-Y. Yang, S. Lindstrom, F. Hormozdiari, E. Eskin, A. L. Price, P. Kraft, and B. Pasaniuc, "Integrating functional data to prioritize causal variants in statistical fine-mapping studies," *PLoS Genet.*, vol. 10, no. 10, p. e1004722, Oct. 2014. 4, 52

[77] H. Kim, K. Y. Lim, J. W. Park, J. Kang, J. K. Won, K. Lee, Y. Shim, C. Park, S. Kim, S. Choi, T. M. Kim, H. Yun, and a. . Park, S", "Sporadic and lynch syndrome-associated mismatch repair-deficient brain tumors." 76

[78] H. S. Kim, J. D. Minna, and M. A. White, "GWAS meets TCGA to illuminate mechanisms of cancer predisposition," *Cell*, vol. 152, no. 3, pp. 387–389, Jan. 2013. 76

[79] A. G. Knudson, "Cancer genetics," *Am. J. Med. Genet.*, vol. 111, no. 1, pp. 96–102, Jul. 2002. 4, 76

[80] D. C. Koboldt, Q. Zhang, D. E. Larson, D. Shen, M. D. McLellan, L. Lin, C. A. Miller, E. R. Mardis, L. Ding, and R. K. Wilson, "VarScan 2: somatic mutation and copy

number alteration discovery in cancer by exome sequencing," *Genome Res.*, vol. 22, no. 3, pp. 568–576, Mar. 2012. 99

[81] A. Koire, P. Katsonis, and O. Lichtarge, "Repurposing germline exams of the cancer genome atlas demands a cautious approach and sample-specific variant filtering," *Pac. Symp. Biocomput.*, vol. 21, pp. 207–218, 2016. 77, 100

[82] P. Kraft and C. A. Haiman, "GWAS identifies a common breast cancer risk allele among BRCA1 carriers," *Nat. Genet.*, vol. 42, no. 10, pp. 819–820, Oct. 2010. 76

[83] M. Lam, C.-Y. Chen, Z. Li, A. R. Martin, J. Bryois, X. Ma, H. Gaspar, M. Ikeda, B. Benyamin, B. C. Brown, R. Liu, W. Zhou, L. Guan, Y. Kamatani, S.-W. Kim, M. Kubo, A. A. A. A. Kusumawardhani, C.-M. Liu, H. Ma, S. Periyasamy, A. Takahashi, Z. Xu, H. Yu, F. Zhu, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Indonesia Schizophrenia Consortium, Genetic REsearch on schizophreniA neTwork-China and the Netherlands (GREAT-CN), W. J. Chen, S. Faraone, S. J. Glatt, L. He, S. E. Hyman, H.-G. Hwu, S. A. McCarroll, B. M. Neale, P. Sklar, D. B. Wildenauer, X. Yu, D. Zhang, B. J. Mowry, J. Lee, P. Holmans, S. Xu, P. F. Sullivan, S. Ripke, M. C. O'Donovan, M. J. Daly, S. Qin, P. Sham, N. Iwata, K. S. Hong, S. G. Schwab, W. Yue, M. Tsuang, J. Liu, X. Ma, R. S. Kahn, Y. Shi, and H. Huang, "Comparative genetic architectures of schizophrenia in east asian and european populations," *Nat. Genet.*, vol. 51, no. 12, pp. 1670–1678, Dec. 2019. 46, 61, 67

[84] C. H. Lee, H. Shi, B. Pasaniuc, E. Eskin, and B. Han, "PLEIO: a method to map and interpret pleiotropic loci with GWAS summary statistics," *Am. J. Hum. Genet.*, vol. 108, no. 1, pp. 36–48, Jan. 2021. 6, 34, 112

[85] Y. Lee, F. Luca, R. Pique-Regi, and X. Wen, "Bayesian Multi-SNP genetic association analysis: Control of FDR and use of summary statistics," *bioRxiv*, p. 316471, May 2018. 59, 61

[86] C. H. Li, S. Haider, Y.-J. Shiah, K. Thai, and P. C. Boutros, "Sex differences in cancer

driver genes and biomarkers," *Cancer Res.*, vol. 78, no. 19, pp. 5527–5537, Oct. 2018. 77, 83, 84

[87] C. H. Li, S. Haider, and P. C. Boutros, "Age influences on the molecular presentation of tumours," *Nat. Commun.*, vol. 13, no. 1, p. 208, Jan. 2022. 77, 83

[88] M. Liu, Y. Jiang, R. Wedow, Y. Li, D. M. Brazel, F. Chen, G. Datta, J. Davila-Velderrain, D. McGuire, C. Tian, X. Zhan, 23andMe Research Team, HUNT All-In Psychiatry, H. Choquet, A. R. Docherty, J. D. Faul, J. R. Foerster, L. G. Fritsche, M. E. Gabrielsen, S. D. Gordon, J. Haessler, J.-J. Hottenga, H. Huang, S.-K. Jang, P. R. Jansen, Y. Ling, R. Mägi, N. Matoba, G. McMahon, A. Mulas, V. Orrù, T. Palviainen, A. Pandit, G. W. Reginsson, A. H. Skogholt, J. A. Smith, A. E. Taylor, C. Turman, G. Willemsen, H. Young, K. A. Young, G. J. M. Zajac, W. Zhao, W. Zhou, G. Bjornsdottir, J. D. Boardman, M. Boehnke, D. I. Boomsma, C. Chen, F. Cucca, G. E. Davies, C. B. Eaton, M. A. Ehringer, T. Esko, E. Fiorillo, N. A. Gillespie, D. F. Gudbjartsson, T. Haller, K. M. Harris, A. C. Heath, J. K. Hewitt, I. B. Hickie, J. E. Hokanson, C. J. Hopfer, D. J. Hunter, W. G. Iacono, E. O. Johnson, Y. Kamatani, S. L. R. Kardia, M. C. Keller, M. Kellis, C. Kooperberg, P. Kraft, K. S. Krauter, M. Laakso, P. A. Lind, A. Loukola, S. M. Lutz, P. A. F. Madden, N. G. Martin, M. McGue, M. B. McQueen, S. E. Medland, A. Metspalu, K. L. Mohlke, J. B. Nielsen, Y. Okada, U. Peters, T. J. C. Polderman, D. Posthuma, A. P. Reiner, J. P. Rice, E. Rimm, R. J. Rose, V. Runarsdottir, M. C. Stallings, A. Stančáková, H. Stefansson, K. K. Thai, H. A. Tindle, T. Tyrfingsson, T. L. Wall, D. R. Weir, C. Weisner, J. B. Whitfield, B. S. Winsvold, J. Yin, L. Zuccolo, L. J. Bierut, K. Hveem, J. J. Lee, M. R. Munafò, N. L. Saccone, C. J. Willer, M. C. Cornelis, S. P. David, D. A. Hinds, E. Jorgenson, J. Kaprio, J. A. Stitzel, K. Stefansson, T. E. Thorgeirsson, G. Abecasis, D. J. Liu, and S. Vrieze, "Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use," *Nat. Genet.*, vol. 51, no. 2, pp. 237–244, Feb. 2019. 89, 105

[89] P. Liu, C. Morrison, L. Wang, D. Xiong, P. Vedell, P. Cui, X. Hua, F. Ding, Y. Lu,

M. James, J. D. Ebben, H. Xu, A. A. Adjei, K. Head, J. W. Andrae, M. R. Tschannen, H. Jacob, J. Pan, Q. Zhang, F. Van den Bergh, H. Xiao, K. C. Lo, J. Patel, T. Richmond, M.-A. Watt, T. Albert, R. Selzer, M. Anderson, J. Wang, Y. Wang, S. Starnes, P. Yang, and M. You, "Identification of somatic mutations in non-small cell lung carcinomas using whole-exome sequencing," *Carcinogenesis*, vol. 33, no. 7, pp. 1270–1276, Jul. 2012. 76

[90] Y. Liu, A. Gusev, Y. J. Heng, L. B. Alexandrov, and P. Kraft, "Somatic mutational profiles and germline polygenic risk scores in human cancer," *Genome Med.*, vol. 14, no. 1, p. 14, Feb. 2022. 77

[91] Z. Liu and X. Lin, "Multiple phenotype association tests using summary statistics in genome-wide association studies," *Biometrics*, vol. 74, no. 1, pp. 165–175, Mar. 2018. 6

[92] A. E. Locke, B. Kahali, S. I. Berndt, A. E. Justice, T. H. Pers, F. R. Day, C. Powell, S. Vedantam, M. L. Buchkovich, J. Yang, D. C. Croteau-Chonka, T. Esko, T. Fall, T. Ferreira, S. Gustafsson, Z. Kutalik, J. Luan, R. Mägi, J. C. Randall, T. W. Winkler, A. R. Wood, T. Workalemahu, J. D. Faul, J. A. Smith, J. H. Zhao, W. Zhao, J. Chen, R. Fehrmann, Å. K. Hedman, J. Karjalainen, E. M. Schmidt, D. Absher, N. Amin, D. Anderson, M. Beekman, J. L. Bolton, J. L. Bragg-Gresham, S. Buyske, A. Demirkan, G. Deng, G. B. Ehret, B. Feenstra, M. F. Feitosa, K. Fischer, A. Goel, J. Gong, A. U. Jackson, S. Kanoni, M. E. Kleber, K. Kristiansson, U. Lim, V. Lotay, M. Mangino, I. M. Leach, C. Medina-Gomez, S. E. Medland, M. A. Nalls, C. D. Palmer, D. Pasko, S. Pechlivanis, M. J. Peters, I. Prokopenko, D. Shungin, A. Stančáková, R. J. Strawbridge, Y. J. Sung, T. Tanaka, A. Teumer, S. Trompet, S. W. van der Laan, J. van Setten, J. V. Van Vliet-Ostaptchouk, Z. Wang, L. Yengo, W. Zhang, A. Isaacs, E. Albrecht, J. Ärnlöv, G. M. Arscott, A. P. Attwood, S. Bandinelli, A. Barrett, I. N. Bas, C. Bellis, A. J. Bennett, C. Berne, R. Blagieva, M. Blüher, S. Böhringer, L. L. Bonnycastle, Y. Böttcher, H. A. Boyd, M. Bruinenberg, I. H. Caspersen, Y.-D. I. Chen, R. Clarke, E. W. Daw, A. J. M. de Craen, G. Delgado, M. Dimitriou, A. S. F. Doney,

N. Eklund, K. Estrada, E. Eury, L. Folkersen, R. M. Fraser, M. E. Garcia, F. Geller, V. Giedraitis, B. Gigante, A. S. Go, A. Golay, A. H. Goodall, S. D. Gordon, M. Gorski, H.-J. Grabe, H. Grallert, T. B. Grammer, J. Gräßler, H. Grönberg, C. J. Groves, G. Gusto, J. Haessler, P. Hall, T. Haller, G. Hallmans, C. A. Hartman, M. Hassinen, C. Hayward, N. L. Heard-Costa, Q. Helmer, C. Hengstenberg, O. Holmen, J.-J. Hottenga, A. L. James, J. M. Jeff, Å. Johansson, J. Jolley, T. Juliusdottir, L. Kinnunen, W. Koenig, M. Koskenvuo, W. Kratzer, J. Laitinen, C. Lamina, K. Leander, N. R. Lee, P. Lichtner, L. Lind, J. Lindström, K. S. Lo, S. Lobbens, R. Lorbeer, Y. Lu, F. Mach, P. K. E. Magnusson, A. Mahajan, W. L. McArdle, S. McLachlan, C. Menni, S. Merger, E. Mihailov, L. Milani, A. Moayyeri, K. L. Monda, M. A. Morken, A. Mulas, G. Müller, M. Müller-Nurasyid, A. W. Musk, R. Nagaraja, M. M. Nöthen, I. M. Nolte, S. Pilz, N. W. Rayner, F. Renstrom, R. Rettig, J. S. Ried, S. Ripke, N. R. Robertson, L. M. Rose, S. Sanna, H. Scharnagl, S. Scholtens, F. R. Schumacher, W. R. Scott, T. Seufferlein, J. Shi, A. V. Smith, J. Smolonska, A. V. Stanton, V. Steinthorsdottir, K. Stirrups, H. M. Stringham, J. Sundström, M. A. Swertz, A. J. Swift, A.-C. Syvänen, S.-T. Tan, B. O. Tayo, B. Thorand, G. Thorleifsson, J. P. Tyrer, H.-W. Uh, L. Vandenput, F. C. Verhulst, S. H. Vermeulen, N. Verweij, J. M. Vonk, L. L. Waite, H. R. Warren, D. Waterworth, M. N. Weedon, L. R. Wilkens, C. Willenborg, T. Wilsgaard, M. K. Wojczynski, A. Wong, A. F. Wright, Q. Zhang, LifeLines Cohort Study, E. P. Brennan, M. Choi, Z. Dastani, A. W. Drong, P. Eriksson, A. Franco-Cereceda, J. R. Gådin, A. G. Gharavi, M. E. Goddard, R. E. Handsaker, J. Huang, F. Karpe, S. Kathiresan, S. Keildson, K. Kiryluk, M. Kubo, J.-Y. Lee, L. Liang, R. P. Lifton, B. Ma, S. A. McCarroll, A. J. McKnight, J. L. Min, M. F. Moffatt, G. W. Montgomery, J. M. Murabito, G. Nicholson, D. R. Nyholt, Y. Okada, J. R. B. Perry, R. Dorajoo, E. Reinmaa, R. M. Salem, N. Sandholm, R. A. Scott, L. Stolk, A. Takahashi, T. Tanaka, F. M. van 't Hooft, A. A. E. Vinkhuyzen, H.-J. Westra, W. Zheng, K. T. Zondervan, ADIPOGen Consortium, AGEN-BMI Working Group, CARDIOGRAMplusC4D Consortium, CKDGen Consortium, GLGC, ICBP, MAGIC Investigators, MuTHER Consortium, MIGen Consortium, PAGE Consortium, ReproGen Consortium, GENIE

Consortium, International Endogene Consortium, A. C. Heath, D. Arveiler, S. J. L. Bakker, J. Beilby, R. N. Bergman, J. Blangero, P. Bovet, H. Campbell, M. J. Caulfield, G. Cesana, A. Chakravarti, D. I. Chasman, P. S. Chines, F. S. Collins, D. C. Crawford, L. A. Cupples, D. Cusi, J. Danesh, U. de Faire, H. M. den Ruijter, A. F. Dominiczak, R. Erbel, J. Erdmann, J. G. Eriksson, M. Farrall, S. B. Felix, E. Ferrannini, J. Ferrières, I. Ford, N. G. Forouhi, T. Forrester, O. H. Franco, R. T. Gansevoort, P. V. Gejman, C. Gieger, O. Gottesman, V. Gudnason, U. Gyllensten, A. S. Hall, T. B. Harris, A. T. Hattersley, A. A. Hicks, L. A. Hindorff, A. D. Hingorani, A. Hofman, G. Homuth, G. K. Hovingh, S. E. Humphries, S. C. Hunt, E. Hyppönen, T. Illig, K. B. Jacobs, M.-R. Jarvelin, K.-H. Jöckel, B. Johansen, P. Jousilahti, J. W. Jukema, A. M. Jula, J. Kaprio, J. J. P. Kastelein, S. M. Keinanen-Kiukaanniemi, L. A. Kiemeney, P. Knekt, J. S. Kooner, C. Kooperberg, P. Kovacs, A. T. Kraja, M. Kumari, J. Kuusisto, T. A. Lakka, C. Langenberg, L. L. Marchand, T. Lehtimäki, V. Lyssenko, S. Männistö, A. Marette, T. C. Matise, C. A. McKenzie, B. McKnight, F. L. Moll, A. D. Morris, A. P. Morris, J. C. Murray, M. Nelis, C. Ohlsson, A. J. Oldehinkel, K. K. Ong, P. A. F. Madden, G. Pasterkamp, J. F. Peden, A. Peters, D. S. Postma, P. P. Pramstaller, J. F. Price, L. Qi, O. T. Raitakari, T. Rankinen, D. C. Rao, T. K. Rice, P. M. Ridker, J. D. Rioux, M. D. Ritchie, I. Rudan, V. Salomaa, N. J. Samani, J. Saramies, M. A. Sarzynski, H. Schunkert, P. E. H. Schwarz, P. Sever, A. R. Shuldiner, J. Sinisalo, R. P. Stolk, K. Strauch, A. Tönjes, D.-A. Trégouët, A. Tremblay, E. Tremoli, J. Virtamo, M.-C. Vohl, U. Völker, G. Waeber, G. Willemsen, J. C. Witteman, M. C. Zillikens, L. S. Adair, P. Amouyel, F. W. Asselbergs, T. L. Assimes, M. Bochud, B. O. Boehm, E. Boerwinkle, S. R. Bornstein, E. P. Bottinger, C. Bouchard, S. Cauchi, J. C. Chambers, S. J. Chanock, R. S. Cooper, P. I. W. de Bakker, G. Dedoussis, L. Ferrucci, P. W. Franks, P. Froguel, L. C. Groop, C. A. Haiman, A. Hamsten, J. Hui, D. J. Hunter, K. Hveem, R. C. Kaplan, M. Kivimaki, D. Kuh, M. Laakso, Y. Liu, N. G. Martin, W. März, M. Melbye, A. Metspalu, S. Moebus, P. B. Munroe, I. Njølstad, B. A. Oostra, C. N. A. Palmer, N. L. Pedersen, M. Perola, L. Pérusse, U. Peters, C. Power, T. Quertermous, R. Rauramaa, F. Rivadeneira, T. E. Saaristo, D. Saleheen,

N. Sattar, E. E. Schadt, D. Schlessinger, P. E. Slagboom, H. Snieder, T. D. Spector, U. Thorsteinsdottir, M. Stumvoll, J. Tuomilehto, A. G. Uitterlinden, M. Uusitupa, P. van der Harst, M. Walker, H. Wallaschofski, N. J. Wareham, H. Watkins, D. R. Weir, H.-E. Wichmann, J. F. Wilson, P. Zanen, I. B. Borecki, P. Deloukas, C. S. Fox, I. M. Heid, J. R. O'Connell, D. P. Strachan, K. Stefansson, C. M. van Duijn, G. R. Abecasis, L. Franke, T. M. Frayling, M. I. McCarthy, P. M. Visscher, A. Scherag, C. J. Willer, M. Boehnke, K. L. Mohlke, C. M. Lindgren, J. S. Beckmann, I. Barroso, K. E. North, E. Ingelsson, J. N. Hirschhorn, R. J. F. Loos, and E. K. Speliotes, "Genetic studies of body mass index yield new insights for obesity biology," *Nature*, vol. 518, no. 7538, pp. 197–206, Feb. 2015. 27, 45

[93] P.-R. Loh, G. Kichaev, S. Gazal, A. P. Schoech, and A. L. Price, "Mixed-model association for biobank-scale datasets," *Nat. Genet.*, vol. 50, no. 7, pp. 906–908, Jun. 2018. 89, 105

[94] J. A. Lozano, F. Hormozdiari, J. W. j. Joo, B. Han, and E. Eskin, "The multivariate normal distribution framework for analyzing association studies," *bioRxiv*, p. 208199, Oct 2017. 66, 67, 71

[95] C.-H. Lu, S.-H. Lee, K.-H. Liu, Y.-S. Hung, C.-H. Wang, Y.-C. Lin, T.-S. Yeh, and W.-C. Chou, "Older age impacts on survival outcome in patients receiving curative surgery for solid cancer," *Asian J. Surg.*, vol. 41, no. 4, pp. 333–340, Jul. 2018. 92, 94

[96] J. MacArthur, E. Bowler, M. Cerezo, L. Gil, P. Hall, E. Hastings, H. Junkins, A. McMahon, A. Milano, J. Morales, Z. M. Pendlington, D. Welter, T. Burdett, L. Hindorff, P. Flicek, F. Cunningham, and H. Parkinson, "The new NHGRI-EBI catalog of published genome-wide association studies (GWAS catalog)," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D896–D901, Jan. 2017. 6

[97] R. Mägi, M. Horikoshi, T. Sofer, A. Mahajan, H. Kitajima, N. Franceschini, M. I. McCarthy, COGENT-Kidney Consortium, T2D-GENES Consortium, and A. P. Morris, "Trans-ethnic meta-regression of genome-wide association studies accounting for

ancestry increases power for discovery and improves fine-mapping resolution," *Hum. Mol. Genet.*, vol. 26, no. 18, pp. 3639–3650, Sep. 2017. 4, 46

[98] J. B. Maller, G. McVean, J. Byrnes, D. Vukcevic, K. Palin, Z. Su, J. M. M. Howson, A. Auton, S. Myers, A. Morris, M. Pirinen, M. A. Brown, P. R. Burton, M. J. Caulfield, A. Compston, M. Farrall, A. S. Hall, A. T. Hattersley, A. V. S. Hill, C. G. Mathew, M. Pembrey, J. Satsangi, M. R. Stratton, J. Worthington, N. Craddock, M. Hurles, W. Ouwehand, M. Parkes, N. Rahman, A. Duncanson, J. A. Todd, D. P. Kwiatkowski, N. J. Samani, S. C. L. Gough, M. I. McCarthy, P. Deloukas, and P. Donnelly, "Bayesian refinement of association signals for 14 loci in 3 common diseases," *Nat. Genet.*, vol. 44, no. 12, pp. 1294–1301, Oct. 2012. 45

[99] A. Marabelle, M. Fakih, J. Lopez, M. Shah, R. Shapira-Frommer, K. Nakagawa, H. C. Chung, H. L. Kindler, J. A. Lopez-Martin, W. H. Miller, Jr, A. Italiano, S. Kao, S. A. Piha-Paul, J.-P. Delord, R. R. McWilliams, D. A. Fabrizio, D. Aurora-Garg, L. Xu, F. Jin, K. Norwood, and Y.-J. Bang, "Association of tumour mutational burden with outcomes in patients with advanced solid tumours treated with pembrolizumab: prospective biomarker analysis of the multicohort, open-label, phase 2 KEYNOTE-158 study," *Lancet Oncol.*, vol. 21, no. 10, pp. 1353–1365, Oct 2020. 80

[100] L. Marcus, L. A. Fashoyin-Aje, M. Donoghue, M. Yuan, L. Rodriguez, P. S. Gallagher, R. Philip, S. Ghosh, M. R. Theoret, J. A. Beaver, R. Pazdur, and S. J. Lemery, "FDA approval summary: Pembrolizumab for the treatment of tumor mutational Burden-High solid tumors," *Clin. Cancer Res.*, vol. 27, no. 17, pp. 4685–4689, Sep. 2021. 80

[101] U. M. Marigorta and A. Navarro, "High trans-ethnic replicability of GWAS results implies common causal variants," *PLoS Genet.*, vol. 9, no. 6, p. e1003566, Jun. 2013. 46, 61, 67

[102] I. Martincorena, K. M. Raine, M. Gerstung, K. J. Dawson, K. Haase, P. Van Loo, H. Davies, M. R. Stratton, and P. J. Campbell, "Universal patterns of selection in cancer and somatic tissues," *Cell*, vol. 173, no. 7, p. 1823, Jun. 2018. 76

[103] N. Mavaddat, A. C. Antoniou, D. F. Easton, and M. Garcia-Closas, "Genetic susceptibility to breast cancer," *Mol. Oncol.*, vol. 4, no. 3, pp. 174–191, Jun. 2010. 76

[104] M. I. McCarthy, G. R. Abecasis, L. R. Cardon, D. B. Goldstein, J. Little, J. P. A. Ioannidis, and J. N. Hirschhorn, "Genome-wide association studies for complex traits: consensus, uncertainty and challenges," *Nat. Rev. Genet.*, vol. 9, no. 5, pp. 356–369, May 2008. 6, 35

[105] S. McCarthy, S. Das, W. Kretzschmar, O. Delaneau, A. R. Wood, A. Teumer, H. M. Kang, C. Fuchsberger, P. Danecek, K. Sharp, Y. Luo, C. Sidore, A. Kwong, N. Timpson, S. Koskinen, S. Vrieze, L. J. Scott, H. Zhang, A. Mahajan, J. Veldink, U. Peters, C. Pato, C. M. van Duijn, C. E. Gillies, I. Gandin, M. Mezzavilla, A. Gilly, M. Cocca, M. Traglia, A. Angius, J. C. Barrett, D. Boomsma, K. Branham, G. Breen, C. M. Brummett, F. Busonero, H. Campbell, A. Chan, S. Chen, E. Chew, F. S. Collins, L. J. Corbin, G. D. Smith, G. Dedoussis, M. Dorr, A.-E. Farmaki, L. Ferrucci, L. Forer, R. M. Fraser, S. Gabriel, S. Levy, L. Groop, T. Harrison, A. Hattersley, O. L. Holmen, K. Hveem, M. Kretzler, J. C. Lee, M. McGue, T. Meitinger, D. Melzer, J. L. Min, K. L. Mohlke, J. B. Vincent, M. Nauck, D. Nickerson, A. Palotie, M. Pato, N. Pirastu, M. McInnis, J. B. Richards, C. Sala, V. Salomaa, D. Schlessinger, S. Schoenherr, P. E. Slagboom, K. Small, T. Spector, D. Stambolian, M. Tuke, J. Tuomilehto, L. H. Van den Berg, W. Van Rheenen, U. Volker, C. Wijmenga, D. Toniolo, E. Zeggini, P. Gasparini, M. G. Sampson, J. F. Wilson, T. Frayling, P. I. W. de Bakker, M. A. Swertz, S. McCarroll, C. Kooperberg, A. Dekker, D. Altshuler, C. Willer, W. Iacono, S. Ripatti, N. Soranzo, K. Walter, A. Swaroop, F. Cucca, C. A. Anderson, R. M. Myers, M. Boehnke, M. I. McCarthy, R. Durbin, and Haplotype Reference Consortium, "A reference panel of 64,976 haplotypes for genotype imputation," *Nat. Genet.*, vol. 48, no. 10, pp. 1279–1283, Oct. 2016. 103

[106] D. J. McGrail, P. G. Pilié, N. U. Rashid, L. Voorwerk, M. Slagter, M. Kok, E. Jonasch, M. Khasraw, A. B. Heimberger, B. Lim, N. T. Ueno, J. K. Litton, R. Ferrarotto, J. T. Chang, S. L. Moulder, and S.-Y. Lin, "High tumor mutation burden fails to predict

immune checkpoint blockade response across all cancer types," *Ann. Oncol.*, vol. 32, no. 5, pp. 661–672, May 2021. 100

[107] J. D. McKay, R. J. Hung, Y. Han, X. Zong, R. Carreras-Torres, D. C. Christiani, N. E. Caporaso, M. Johansson, X. Xiao, Y. Li, J. Byun, A. Dunning, K. A. Pooley, D. C. Qian, X. Ji, G. Liu, M. N. Timofeeva, S. E. Bojesen, X. Wu, L. Le Marchand, D. Albanes, H. Bickeböller, M. C. Aldrich, W. S. Bush, A. Tardon, G. Rennert, M. D. Teare, J. K. Field, L. A. Kiemeney, P. Lazarus, A. Haugen, S. Lam, M. B. Schabath, A. S. Andrew, H. Shen, Y.-C. Hong, J.-M. Yuan, P. A. Bertazzi, A. C. Pesatori, Y. Ye, N. Diao, L. Su, R. Zhang, Y. Brhane, N. Leighl, J. S. Johansen, A. Mellemgaard, W. Saliba, C. A. Haiman, L. R. Wilkens, A. Fernandez-Somoano, G. Fernandez-Tardon, H. F. M. van der Heijden, J. H. Kim, J. Dai, Z. Hu, M. P. A. Davies, M. W. Marcus, H. Brunnström, J. Manjer, O. Melander, D. C. Muller, K. Overvad, A. Trichopoulou, R. Tumino, J. A. Doherty, M. P. Barnett, C. Chen, G. E. Goodman, A. Cox, F. Taylor, P. Woll, I. Brüske, H.-E. Wichmann, J. Manz, T. R. Muley, A. Risch, A. Rosenberger, K. Grankvist, M. Johansson, F. A. Shepherd, M.-S. Tsao, S. M. Arnold, E. B. Haura, C. Bolca, I. Holcatova, V. Janout, M. Kontic, J. Lissowska, A. Mukeria, S. Ognjanovic, T. M. Orlowski, G. Scelo, B. Swiatkowska, D. Zaridze, P. Bakke, V. Skaug, S. Zienolddiny, E. J. Duell, L. M. Butler, W.-P. Koh, Y.-T. Gao, R. S. Houlston, J. McLaughlin, V. L. Stevens, P. Joubert, M. Lamontagne, D. C. Nickle, M. Obeidat, W. Timens, B. Zhu, L. Song, L. Kachuri, M. S. Artigas, M. D. Tobin, L. V. Wain, SpiroMeta Consortium, T. Rafnar, T. E. Thorgeirsson, G. W. Reginsson, K. Stefansson, D. B. Hancock, L. J. Bierut, M. R. Spitz, N. C. Gaddis, S. M. Lutz, F. Gu, E. O. Johnson, A. Kamal, C. Pikielny, D. Zhu, S. Lindströem, X. Jiang, R. F. Tyndale, G. Chenevix-Trench, J. Beesley, Y. Bossé, S. Chanock, P. Brennan, M. T. Landi, and C. I. Amos, "Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes," *Nat. Genet.*, vol. 49, no. 7, pp. 1126–1132, Jul. 2017. 89, 105

[108] B. S. Melin, J. S. Barnholtz-Sloan, M. R. Wrensch, C. Johansen, D. Il'yasova, B. Kin-

nersley, Q. T. Ostrom, K. Labreche, Y. Chen, G. Armstrong, Y. Liu, J. E. Eckel-Passow, P. A. Decker, M. Labussière, A. Idbaih, K. Hoang-Xuan, A.-L. Di Stefano, K. Mokhtari, J.-Y. Delattre, P. Broderick, P. Galan, K. Gousias, J. Schramm, M. J. Schoemaker, S. J. Fleming, S. Herms, S. Heilmann, M. M. Nöthen, H.-E. Wichmann, S. Schreiber, A. Swerdlow, M. Lathrop, M. Simon, M. Sanson, U. Andersson, P. Rajaraman, S. Chanock, M. Linet, Z. Wang, M. Yeager, GliomaScan Consortium, J. K. Wiencke, H. Hansen, L. McCoy, T. Rice, M. L. Kosel, H. Sicotte, C. I. Amos, J. L. Bernstein, F. Davis, D. Lachance, C. Lau, R. T. Merrell, J. Shildkraut, F. Ali-Osman, S. Sadetzki, M. Scheurer, S. Shete, R. K. Lai, E. B. Claus, S. H. Olson, R. B. Jenkins, R. S. Houlston, and M. L. Bondy, "Genome-wide association study of glioma subtypes identifies specific differences in genetic susceptibility to glioblastoma and non-glioblastoma tumors," *Nat. Genet.*, vol. 49, no. 5, pp. 789–794, May 2017. 89, 105

[109] A. P. Morris, "Transethnic meta-analysis of genomewide association studies," *Genet. Epidemiol.*, vol. 35, no. 8, pp. 809–822, Dec. 2011. 4, 46

[110] B. Neale, L. Abbott, V. Anttila, K. Aragam, A. Baumann, J. Bloom, S. Bryant, C. Churchhouse, J. Cole, M. J. Daly, R. Damian, A. Ganna, J. Goldstein, M. Haas, J. Hirschhorn, D. Howrigan, E. Jones, D. King, and R. Munshi, "Uk biobank summary statistics (round 1)," http://www.nealelab.is/blog/2017/7/19/rapid-gwas-of-thousands-of-phenotypes-for-337000-samples-in-the-uk-biobank, 2017. 7, 31, 42

[111] B. Neale Lab, "Uk biobank summary statistics (round 2)," http://www.nealelab.is/uk-biobank, 2018. 7, 17, 24, 42, 49, 51

[112] P. J. Newcombe, D. V. Conti, and S. Richardson, "JAM: A scalable bayesian framework for joint analysis of marginal SNP effects," *Genet. Epidemiol.*, vol. 40, no. 3, pp. 188–201, Apr. 2016. 46

[113] J. Nishino, H. Ochi, Y. Kochi, T. Tsunoda, and S. Matsui, "Sample size for successful

Genome-Wide association study of major depressive disorder," *Front. Genet.*, vol. 9, p. 227, Jun. 2018. 3, 6

[114] B. Niu, K. Ye, Q. Zhang, C. Lu, M. Xie, M. D. McLellan, M. C. Wendl, and L. Ding, "MSIsensor: microsatellite instability detection using paired tumor-normal sequence data," *Bioinformatics*, vol. 30, no. 7, pp. 1015–1016, Apr. 2014. 103

[115] P. F. O'Reilly, C. J. Hoggart, Y. Pomyen, F. C. F. Calboli, P. Elliott, M.-R. Jarvelin, and L. J. M. Coin, "MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS," *PLoS One*, vol. 7, no. 5, p. e34861, May 2012. 6

[116] C. Palmer and I. Pe'er, "Statistical correction of the winner's curse explains replication variability in quantitative trait genome-wide association studies," *PLoS Genet.*, vol. 13, no. 7, p. e1006916, Jul. 2017. 29, 33

[117] O. A. Panagiotou, J. P. A. Ioannidis, and Genome-Wide Significance Project, "What should the genome-wide significance threshold be? empirical replication of borderline genetic associations," *Int. J. Epidemiol.*, vol. 41, no. 1, pp. 273–286, Feb. 2012. 74

[118] B. Pasaniuc and A. L. Price, "Dissecting the genetics of complex traits using summary association statistics," *Nat. Rev. Genet.*, vol. 18, no. 2, pp. 117–127, Feb. 2017. 8

[119] I. Pe'er, R. Yelensky, D. Altshuler, and M. J. Daly, "Estimation of the multiple testing burden for genomewide association studies of nearly all common variants," *Genet. Epidemiol.*, vol. 32, no. 4, pp. 381–385, May 2008. 35

[120] R. E. Peterson, K. Kuchenbaecker, R. K. Walters, C.-Y. Chen, A. B. Popejoy, S. Periyasamy, M. Lam, C. Iyegbe, R. J. Strawbridge, L. Brick, C. E. Carey, A. R. Martin, J. L. Meyers, J. Su, J. Chen, A. C. Edwards, A. Kalungi, N. Koen, L. Majara, E. Schwarz, J. W. Smoller, E. A. Stahl, P. F. Sullivan, E. Vassos, B. Mowry, M. L. Prieto, A. Cuellar-Barboza, T. B. Bigdeli, H. J. Edenberg, H. Huang, and L. E. Duncan, "Genome-wide association studies in ancestrally diverse populations: Opportunities,

methods, pitfalls, and recommendations," *Cell*, vol. 179, no. 3, pp. 589–603, Oct. 2019. 46, 61, 67

[121] C. M. Phelan, K. B. Kuchenbaecker, J. P. Tyrer, S. P. Kar, K. Lawrenson, S. J. Winham, J. Dennis, A. Pirie, M. J. Riggan, G. Chornokur, M. A. Earp, P. C. Lyra, Jr, J. M. Lee, S. Coetzee, J. Beesley, L. McGuffog, P. Soucy, E. Dicks, A. Lee, D. Barrowdale, J. Lecarpentier, G. Leslie, C. M. Aalfs, K. K. H. Aben, M. Adams, J. Adlard, I. L. Andrulis, H. Anton-Culver, N. Antonenkova, AOCS study group, G. Aravantinos, N. Arnold, B. K. Arun, B. Arver, J. Azzollini, J. Balmaña, S. N. Banerjee, L. Barjhoux, R. B. Barkardottir, Y. Bean, M. W. Beckmann, A. Beeghly-Fadiel, J. Benitez, M. Bermisheva, M. Q. Bernardini, M. J. Birrer, L. Bjorge, A. Black, K. Blankstein, M. J. Blok, C. Bodelon, N. Bogdanova, A. Bojesen, B. Bonanni, Å. Borg, A. R. Bradbury, J. D. Brenton, C. Brewer, L. Brinton, P. Broberg, A. Brooks-Wilson, F. Bruinsma, J. Brunet, B. Buecher, R. Butzow, S. S. Buys, T. Caldes, M. A. Caligo, I. Campbell, R. Cannioto, M. E. Carney, T. Cescon, S. B. Chan, J. Chang-Claude, S. Chanock, X. Q. Chen, Y.-E. Chiew, J. Chiquette, W. K. Chung, K. B. M. Claes, T. Conner, L. S. Cook, J. Cook, D. W. Cramer, J. M. Cunningham, A. A. D'Aloisio, M. B. Daly, F. Damiola, S. D. Damirovna, A. Dansonka-Mieszkowska, F. Dao, R. Davidson, A. DeFazio, C. Delnatte, K. F. Doheny, O. Diez, Y. C. Ding, J. A. Doherty, S. M. Domchek, C. M. Dorfling, T. Dörk, L. Dossus, M. Duran, M. Dürst, B. Dworniczak, D. Eccles, T. Edwards, R. Eeles, U. Eilber, B. Ejlertsen, A. B. Ekici, S. Ellis, M. Elvira, EMBRACE Study, K. H. Eng, C. Engel, D. G. Evans, P. A. Fasching, S. Ferguson, S. F. Ferrer, J. M. Flanagan, Z. C. Fogarty, R. T. Fortner, F. Fostira, W. D. Foulkes, G. Fountzilas, B. L. Fridley, T. M. Friebel, E. Friedman, D. Frost, P. A. Ganz, J. Garber, M. J. García, V. Garcia-Barberan, A. Gehrig, GEMO Study Collaborators, A. Gentry-Maharaj, A.-M. Gerdes, G. G. Giles, R. Glasspool, G. Glendon, A. K. Godwin, D. E. Goldgar, T. Goranova, M. Gore, M. H. Greene, J. Gronwald, S. Gruber, E. Hahnen, C. A. Haiman, N. Håkansson, U. Hamann, T. V. O. Hansen, P. A. Harrington, H. R. Harris, J. Hauke, HEBON Study, A. Hein, A. Hen-

derson, M. A. T. Hildebrandt, P. Hillemanns, S. Hodgson, C. K. Høgdall, E. Høgdall, F. B. L. Hogervorst, H. Holland, M. J. Hooning, K. Hosking, R.-Y. Huang, P. J. Hulick, J. Hung, D. J. Hunter, D. G. Huntsman, T. Huzarski, E. N. Imyanitov, C. Isaacs, E. S. Iversen, L. Izatt, A. Izquierdo, A. Jakubowska, P. James, R. Janavicius, M. Jernetz, A. Jensen, U. B. Jensen, E. M. John, S. Johnatty, M. E. Jones, P. Kannisto, B. Y. Karlan, A. Karnezis, K. Kast, KConFab Investigators, C. J. Kennedy, E. Khusnutdinova, L. A. Kiemeney, J. I. Kiiski, S.-W. Kim, S. K. Kjaer, M. Köbel, R. K. Kopperud, T. A. Kruse, J. Kupryjanczyk, A. Kwong, Y. Laitman, D. Lambrechts, N. Larrañaga, M. C. Larson, C. Lazaro, N. D. Le, L. Le Marchand, J. W. Lee, S. B. Lele, A. Leminen, D. Leroux, J. Lester, F. Lesueur, D. A. Levine, D. Liang, C. Liebrich, J. Lilyquist, L. Lipworth, J. Lissowska, K. H. Lu, J. Lubiński, C. Luccarini, L. Lundvall, P. L. Mai, G. Mendoza-Fandiño, S. Manoukian, L. F. A. G. Massuger, T. May, S. Mazoyer, J. N. McAlpine, V. McGuire, J. R. McLaughlin, I. McNeish, H. Meijers-Heijboer, A. Meindl, U. Menon, A. R. Mensenkamp, M. A. Merritt, R. L. Milne, G. Mitchell, F. Modugno, J. Moes-Sosnowska, M. Moffitt, M. Montagna, K. B. Moysich, A. M. Mulligan, J. Musinsky, K. L. Nathanson, L. Nedergaard, R. B. Ness, S. L. Neuhausen, H. Nevanlinna, D. Niederacher, R. L. Nussbaum, K. Odunsi, E. Olah, O. I. Olopade, H. Olsson, C. Olswold, D. M. O'Malley, K.-R. Ong, N. C. Onland-Moret, OPAL study group, N. Orr, S. Orsulic, A. Osorio, D. Palli, L. Papi, T.-W. Park-Simon, J. Paul, C. L. Pearce, I. S. Pedersen, P. H. M. Peeters, B. Peissel, A. Peixoto, T. Pejovic, L. M. Pelttari, J. B. Permuth, P. Peterlongo, L. Pezzani, G. Pfeiler, K.-A. Phillips, M. Piedmonte, M. C. Pike, A. M. Piskorz, S. R. Poblete, T. Pocza, E. M. Poole, B. Poppe, M. E. Porteous, F. Prieur, D. Prokofyeva, E. Pugh, M. A. Pujana, P. Pujol, P. Radice, J. Rantala, C. Rappaport-Fuerhauser, G. Rennert, K. Rhiem, P. Rice, A. Richardson, M. Robson, G. C. Rodriguez, C. Rodríguez-Antona, J. Romm, M. A. Rookus, M. A. Rossing, J. H. Rothstein, A. Rudolph, I. B. Runnebaum, H. B. Salvesen, D. P. Sandler, M. J. Schoemaker, L. Senter, V. W. Setiawan, G. Severi, P. Sharma, T. Shelford, N. Siddiqui, L. E. Side, W. Sieh, C. F. Singer, H. Sobol, H. Song, M. C. Southey, A. B. Spurdle, Z. Stadler, D. Steinemann, D. Stoppa-Lyonnet, L. E. Sucheston-Campbell,

G. Sukiennicki, R. Sutphen, C. Sutter, A. J. Swerdlow, C. I. Szabo, L. Szafron, Y. Y. Tan, J. A. Taylor, M.-K. Tea, M. R. Teixeira, S.-H. Teo, K. L. Terry, P. J. Thompson, L. C. V. Thomsen, D. L. Thull, L. Tihomirova, A. V. Tinker, M. Tischkowitz, S. Tognazzo, A. E. Toland, A. Tone, B. Trabert, R. C. Travis, A. Trichopoulou, N. Tung, S. S. Tworoger, A. M. van Altena, D. Van Den Berg, A. H. van der Hout, R. B. van der Luijt, M. Van Heetvelde, E. Van Nieuwenhuysen, E. J. van Rensburg, A. Vanderstichele, R. Varon-Mateeva, A. Vega, D. V. Edwards, I. Vergote, R. A. Vierkant, J. Vijai, A. Vratimos, L. Walker, C. Walsh, D. Wand, S. Wang-Gohrke, B. Wappenschmidt, P. M. Webb, C. R. Weinberg, J. N. Weitzel, N. Wentzensen, A. S. Whittemore, J. T. Wijnen, L. R. Wilkens, A. Wolk, M. Woo, X. Wu, A. H. Wu, H. Yang, D. Yannoukakos, A. Ziogas, K. K. Zorn, S. A. Narod, D. F. Easton, C. I. Amos, J. M. Schildkraut, S. J. Ramus, L. Ottini, M. T. Goodman, S. K. Park, L. E. Kelemen, H. A. Risch, M. Thomassen, K. Offit, J. Simard, R. K. Schmutzler, D. Hazelett, A. N. Monteiro, F. J. Couch, A. Berchuck, G. Chenevix-Trench, E. L. Goode, T. A. Sellers, S. A. Gayther, A. C. Antoniou, and P. D. P. Pharoah, "Identification of 12 new susceptibility loci for different histotypes of epithelial ovarian cancer," *Nat. Genet.*, vol. 49, no. 5, pp. 680–691, May 2017. 89, 105

[122] E. D. Pleasance, R. K. Cheetham, P. J. Stephens, D. J. McBride, S. J. Humphray, C. D. Greenman, I. Varela, M.-L. Lin, G. R. Ordóñez, G. R. Bignell, K. Ye, J. Alipaz, M. J. Bauer, D. Beare, A. Butler, R. J. Carter, L. Chen, A. J. Cox, S. Edkins, P. I. Kokko-Gonzales, N. A. Gormley, R. J. Grocock, C. D. Haudenschild, M. M. Hims, T. James, M. Jia, Z. Kingsbury, C. Leroy, J. Marshall, A. Menzies, L. J. Mudie, Z. Ning, T. Royce, O. B. Schulz-Trieglaff, A. Spiridou, L. A. Stebbings, L. Szajkowski, J. Teague, D. Williamson, L. Chin, M. T. Ross, P. J. Campbell, D. R. Bentley, P. A. Futreal, and M. R. Stratton, "A comprehensive catalogue of somatic mutations from a human cancer genome," *Nature*, vol. 463, no. 7278, pp. 191–196, Jan. 2010. 4, 76

[123] B. A. Ponder, "Cancer genetics," *Nature*, vol. 411, no. 6835, pp. 336–341, May 2001. 4, 76

[124] E. Porta-Pardo, R. Sayaman, E. Ziv, and A. Valencia, "The landscape of interactions between cancer polygenic risk scores and somatic alterations in cancer cells," *bioRxiv*, p. 2020.09.28.316851, Sep. 2020. 77

[125] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, and P. C. Sham, "PLINK: a tool set for whole-genome association and population-based linkage analyses," *Am. J. Hum. Genet.*, vol. 81, no. 3, pp. 559–575, Sep. 2007. 101, 104, 105

[126] G. Qi and N. Chatterjee, "Heritability informed power optimization (HIPO) leads to enhanced detection of genetic associations across multiple traits," *PLoS Genet.*, vol. 14, no. 10, p. e1007549, Oct. 2018. 6, 7, 15, 24

[127] R. Rasnic, N. Brandes, O. Zuk, and M. Linial, "Substantial batch effects in TCGA exome sequences undermine pan-cancer analysis of germline variants," *BMC Cancer*, vol. 19, no. 1, p. 783, Aug. 2019. 77, 100

[128] E. Rheinbay, M. M. Nielsen, F. Abascal, J. A. Wala, O. Shapira, G. Tiao, H. Hornshøj, J. M. Hess, R. I. Juul, Z. Lin, L. Feuerbach, R. Sabarinathan, T. Madsen, J. Kim, L. Mularoni, S. Shuai, A. Lanzós, C. Herrmann, Y. E. Maruvka, C. Shen, S. B. Amin, P. Bandopadhayay, J. Bertl, K. A. Boroevich, J. Busanovich, J. Carlevaro-Fita, D. Chakravarty, C. W. Y. Chan, D. Craft, P. Dhingra, K. Diamanti, N. A. Fonseca, A. Gonzalez-Perez, Q. Guo, M. P. Hamilton, N. J. Haradhvala, C. Hong, K. Isaev, T. A. Johnson, M. Juul, A. Kahles, A. Kahraman, Y. Kim, J. Komorowski, K. Kumar, S. Kumar, D. Lee, K.-V. Lehmann, Y. Li, E. M. Liu, L. Lochovsky, K. Park, O. Pich, N. D. Roberts, G. Saksena, S. E. Schumacher, N. Sidiropoulos, L. Sieverling, N. Sinnott-Armstrong, C. Stewart, D. Tamborero, J. M. C. Tubio, H. M. Umer, L. Uusküla-Reimand, C. Wadelius, L. Wadi, X. Yao, C.-Z. Zhang, J. Zhang, J. E. Haber, A. Hobolth, M. Imielinski, M. Kellis, M. S. Lawrence, C. von Mering, H. Nakagawa, B. J. Raphael, M. A. Rubin, C. Sander, L. D. Stein, J. M. Stuart, T. Tsunoda, D. A. Wheeler, R. Johnson, J. Reimand, M. Gerstein, E. Khurana, P. J. Camp-

bell, N. López-Bigas, PCAWG Drivers and Functional Interpretation Working Group, PCAWG Structural Variation Working Group, J. Weischenfeldt, R. Beroukhim, I. Martincorena, J. S. Pedersen, G. Getz, and PCAWG Consortium, "Analyses of non-coding somatic drivers in 2,658 cancer whole genomes," *Nature*, vol. 578, no. 7793, pp. 102–111, Feb. 2020. 76

[129] M. Riihimäki, H. Thomsen, A. Hemminki, K. Sundquist, and K. Hemminki, "Comparison of survival of patients with metastases from known versus unknown primaries: survival in metastatic cancer," *BMC Cancer*, vol. 13, p. 36, Jan. 2013. 92, 94

[130] P. Riviere, A. M. Goodman, R. Okamura, D. A. Barkauskas, T. J. Whitchurch, S. Lee, N. Khalid, R. Collier, M. Mareboina, G. M. Frampton, D. Fabrizio, A. B. Sharabi, S. Kato, and R. Kurzrock, "High tumor mutational burden correlates with longer survival in Immunotherapy-Naïve patients with diverse cancers," *Mol. Cancer Ther.*, vol. 19, no. 10, pp. 2139–2145, Oct. 2020. 76

[131] S. Sakaue, M. Kanai, Y. Tanigawa, J. Karjalainen, M. Kurki, S. Koshiba, A. Narita, T. Konuma, K. Yamamoto, M. Akiyama, K. Ishigaki, A. Suzuki, K. Suzuki, W. Obara, K. Yamaji, K. Takahashi, S. Asai, Y. Takahashi, T. Suzuki, N. Shinozaki, H. Yamaguchi, S. Minami, S. Murayama, K. Yoshimori, S. Nagayama, D. Obata, M. Higashiyama, A. Masumoto, Y. Koretsune, FinnGen, K. Ito, C. Terao, T. Yamauchi, I. Komuro, T. Kadowaki, G. Tamiya, M. Yamamoto, Y. Nakamura, M. Kubo, Y. Murakami, K. Yamamoto, Y. Kamatani, A. Palotie, M. A. Rivas, M. J. Daly, K. Matsuda, and Y. Okada, "A cross-population atlas of genetic associations for 220 human phenotypes," *Nat. Genet.*, vol. 53, no. 10, pp. 1415–1424, Oct. 2021. 3, 51

[132] R. M. Samstein, C.-H. Lee, A. N. Shoushtari, M. D. Hellmann, R. Shen, Y. Y. Janjigian, D. A. Barron, A. Zehir, E. J. Jordan, A. Omuro, T. J. Kaley, S. M. Kendall, R. J. Motzer, A. A. Hakimi, M. H. Voss, P. Russo, J. Rosenberg, G. Iyer, B. H. Bochner, D. F. Bajorin, H. A. Al-Ahmadie, J. E. Chaft, C. M. Rudin, G. J. Riely, S. Baxi, A. L. Ho, R. J. Wong, D. G. Pfister, J. D. Wolchok, C. A. Barker, P. H.

Gutin, C. W. Brennan, V. Tabar, I. K. Mellinghoff, L. M. DeAngelis, C. E. Ariyan, N. Lee, W. D. Tap, M. M. Gounder, S. P. D'Angelo, L. Saltz, Z. K. Stadler, H. I. Scher, J. Baselga, P. Razavi, C. A. Klebanoff, R. Yaeger, N. H. Segal, G. Y. Ku, R. P. DeMatteo, M. Ladanyi, N. A. Rizvi, M. F. Berger, N. Riaz, D. B. Solit, T. A. Chan, and L. G. T. Morris, "Tumor mutational load predicts survival after immunotherapy across multiple cancer types," *Nat. Genet.*, vol. 51, no. 2, pp. 202–206, Feb 2019. 76, 92

[133] G. Scelo, M. P. Purdue, K. M. Brown, M. Johansson, Z. Wang, J. E. Eckel-Passow, Y. Ye, J. N. Hofmann, J. Choi, M. Foll, V. Gaborieau, M. J. Machiela, L. M. Colli, P. Li, J. N. Sampson, B. Abedi-Ardekani, C. Besse, H. Blanche, A. Boland, L. Burdette, A. Chabrier, G. Durand, F. Le Calvez-Kelm, E. Prokhortchouk, N. Robinot, K. G. Skryabin, M. B. Wozniak, M. Yeager, G. Basta-Jovanovic, Z. Dzamic, L. Foretova, I. Holcatova, V. Janout, D. Mates, A. Mukeriya, S. Rascu, D. Zaridze, V. Bencko, C. Cybulski, E. Fabianova, V. Jinga, J. Lissowska, J. Lubinski, M. Navratilova, P. Rudnai, N. Szeszenia-Dabrowska, S. Benhamou, G. Cancel-Tassin, O. Cussenot, L. Baglietto, H. Boeing, K.-T. Khaw, E. Weiderpass, B. Ljungberg, R. T. Sitaram, F. Bruinsma, S. J. Jordan, G. Severi, I. Winship, K. Hveem, L. J. Vatten, T. Fletcher, K. Koppova, S. C. Larsson, A. Wolk, R. E. Banks, P. J. Selby, D. F. Easton, P. Pharoah, G. Andreotti, L. E. B. Freeman, S. Koutros, D. Albanes, S. Männistö, S. Weinstein, P. E. Clark, T. L. Edwards, L. Lipworth, S. M. Gapstur, V. L. Stevens, H. Carol, M. L. Freedman, M. M. Pomerantz, E. Cho, P. Kraft, M. A. Preston, K. M. Wilson, J. Michael Gaziano, H. D. Sesso, A. Black, N. D. Freedman, W.-Y. Huang, J. G. Anema, R. J. Kahnoski, B. R. Lane, S. L. Noyes, D. Petillo, B. T. Teh, U. Peters, E. White, G. L. Anderson, L. Johnson, J. Luo, J. Buring, I.-M. Lee, W.-H. Chow, L. E. Moore, C. Wood, T. Eisen, M. Henrion, J. Larkin, P. Barman, B. C. Leibovich, T. K. Choueiri, G. Mark Lathrop, N. Rothman, J.-F. Deleuze, J. D. McKay, A. S. Parker, X. Wu, R. S. Houlston, P. Brennan, and S. J. Chanock, "Genome-wide association study identifies multiple risk loci for renal cell carcinoma," *Nat. Commun.*,

vol. 8, p. 15724, Jun. 2017. 89, 105

[134] D. J. Schaid, W. Chen, and N. B. Larson, "From genome-wide associations to candidate causal variants by statistical fine-mapping," *Nat. Rev. Genet.*, vol. 19, no. 8, pp. 491–504, Aug. 2018. 45

[135] D. Schnidrig, S. Turajlic, and K. Litchfield, "Tumour mutational burden: primary versus metastatic tissue creates systematic bias," *Immunooncol Technol*, vol. 4, pp. 8–14, Dec. 2019. 77, 84

[136] F. R. Schumacher, A. A. Al Olama, S. I. Berndt, S. Benlloch, M. Ahmed, E. J. Saunders, T. Dadaev, D. Leongamornlert, E. Anokian, C. Cieza-Borrella, C. Goh, M. N. Brook, X. Sheng, L. Fachal, J. Dennis, J. Tyrer, K. Muir, A. Lophatananon, V. L. Stevens, S. M. Gapstur, B. D. Carter, C. M. Tangen, P. J. Goodman, I. M. Thompson, Jr, J. Batra, S. Chambers, L. Moya, J. Clements, L. Horvath, W. Tilley, G. P. Risbridger, H. Gronberg, M. Aly, T. Nordström, P. Pharoah, N. Pashayan, J. Schleutker, T. L. J. Tammela, C. Sipeky, A. Auvinen, D. Albanes, S. Weinstein, A. Wolk, N. Håkansson, C. M. L. West, A. M. Dunning, N. Burnet, L. A. Mucci, E. Giovannucci, G. L. Andriole, O. Cussenot, G. Cancel-Tassin, S. Koutros, L. E. Beane Freeman, K. D. Sorensen, T. F. Orntoft, M. Borre, L. Maehle, E. M. Grindedal, D. E. Neal, J. L. Donovan, F. C. Hamdy, R. M. Martin, R. C. Travis, T. J. Key, R. J. Hamilton, N. E. Fleshner, A. Finelli, S. A. Ingles, M. C. Stern, B. S. Rosenstein, S. L. Kerns, H. Ostrer, Y.-J. Lu, H.-W. Zhang, N. Feng, X. Mao, X. Guo, G. Wang, Z. Sun, G. G. Giles, M. C. Southey, R. J. MacInnis, L. M. FitzGerald, A. S. Kibel, B. F. Drake, A. Vega, A. Gómez-Caamaño, R. Szulkin, M. Eklund, M. Kogevinas, J. Llorca, G. Castaño-Vinyals, K. L. Penney, M. Stampfer, J. Y. Park, T. A. Sellers, H.-Y. Lin, J. L. Stanford, C. Cybulski, D. Wokolorczyk, J. Lubinski, E. A. Ostrander, M. S. Geybels, B. G. Nordestgaard, S. F. Nielsen, M. Weischer, R. Bisbjerg, M. A. Røder, P. Iversen, H. Brenner, K. Cuk, B. Holleczek, C. Maier, M. Luedeke, T. Schnoeller, J. Kim, C. J. Logothetis, E. M. John, M. R. Teixeira, P. Paulo, M. Cardoso, S. L. Neuhausen, L. Steele, Y. C. Ding, K. De Ruyck, G. De Meerleer, P. Ost, A. Razack,

J. Lim, S.-H. Teo, D. W. Lin, L. F. Newcomb, D. Lessel, M. Gamulin, T. Kulis, R. Kaneva, N. Usmani, S. Singhal, C. Slavov, V. Mitev, M. Parliament, F. Claessens, S. Joniau, T. Van den Broeck, S. Larkin, P. A. Townsend, C. Aukim-Hastie, M. Gago-Dominguez, J. E. Castelao, M. E. Martinez, M. J. Roobol, G. Jenster, R. H. N. van Schaik, F. Menegaux, T. Truong, Y. A. Koudou, J. Xu, K.-T. Khaw, L. Cannon-Albright, H. Pandha, A. Michael, S. N. Thibodeau, S. K. McDonnell, D. J. Schaid, S. Lindstrom, C. Turman, J. Ma, D. J. Hunter, E. Riboli, A. Siddiq, F. Canzian, L. N. Kolonel, L. Le Marchand, R. N. Hoover, M. J. Machiela, Z. Cui, P. Kraft, C. I. Amos, D. V. Conti, D. F. Easton, F. Wiklund, S. J. Chanock, B. E. Henderson, Z. Kote-Jarai, C. A. Haiman, R. A. Eeles, Profile Study, Australian Prostate Cancer BioResource (APCB), IMPACT Study, Canary PASS Investigators, Breast and Prostate Cancer Cohort Consortium (BPC3), PRACTICAL (Prostate Cancer Association Group to Investigate Cancer-Associated Alterations in the Genome) Consortium, Cancer of the Prostate in Sweden (CAPS), Prostate Cancer Genome-wide Association Study of Uncommon Susceptibility Loci (PEGASUS), and Genetic Associations and Mechanisms in Oncology (GAME-ON)/Elucidating Loci Involved in Prostate Cancer Susceptibility (ELLIPSE) Consortium, "Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci," *Nat. Genet.*, vol. 50, no. 7, pp. 928–936, Jul. 2018. 89, 105

[137] M. R. Schwartz, L. Luo, and M. Berwick, "Sex differences in melanoma," *Curr Epidemiol Rep*, vol. 6, no. 2, pp. 112–118, Jun. 2019. 83

[138] D. Sha, Z. Jin, J. Budczies, K. Kluck, A. Stenzinger, and F. A. Sinicrope, "Tumor mutational burden as a predictive biomarker in solid tumors," *Cancer Discov.*, vol. 10, no. 12, pp. 1808–1825, Dec. 2020. 76

[139] H. Shi, G. Kichaev, and B. Pasaniuc, "Contrasting the genetic architecture of 30 complex traits from summary association data," *Am. J. Hum. Genet.*, vol. 99, no. 1, pp. 139–153, Jul. 2016. 34, 111

[140] H. Shi, N. Mancuso, S. Spendlove, and B. Pasaniuc, "Local genetic correlation gives insights into the shared genetic architecture of complex traits," *Am. J. Hum. Genet.*, vol. 101, no. 5, pp. 737–751, Nov. 2017. 34, 111

[141] A. Snyder, V. Makarov, T. Merghoub, J. Yuan, J. M. Zaretsky, A. Desrichard, L. A. Walsh, M. A. Postow, P. Wong, T. S. Ho, T. J. Hollmann, C. Bruggeman, K. Kannan, Y. Li, C. Elipenahli, C. Liu, C. T. Harbison, L. Wang, A. Ribas, J. D. Wolchok, and T. A. Chan, "Genetic basis for clinical response to CTLA-4 blockade in melanoma," *N. Engl. J. Med.*, vol. 371, no. 23, pp. 2189–2199, Dec. 2014. 76

[142] P. W. Sperduto, S. Mesko, J. Li, D. Cagney, A. Aizer, N. U. Lin, E. Nesbit, T. J. Kruser, J. Chan, S. Braunstein, J. Lee, J. P. Kirkpatrick, W. Breen, P. D. Brown, D. Shi, H. A. Shih, H. Soliman, A. Sahgal, R. Shanley, W. A. Sperduto, E. Lou, A. Everett, D. H. Boggs, L. Masucci, D. Roberge, J. Remick, K. Plichta, J. M. Buatti, S. Jain, L. E. Gaspar, C.-C. Wu, T. J. C. Wang, J. Bryant, M. Chuong, Y. An, V. Chiang, T. Nakano, H. Aoyama, and M. P. Mehta, "Survival in patients with brain metastases: Summary report on the updated Diagnosis-Specific graded prognostic assessment and definition of the eligibility quotient," *J. Clin. Orthod.*, vol. 38, no. 32, pp. 3773–3784, Nov. 2020. 92

[143] L. P. Stabile, V. Kumar, A. Gaither-Davis, E. H. Huang, F. P. Vendetti, P. Devadassan, S. Dacic, R. Bao, R. A. Steinman, T. F. Burns, and C. J. Bakkenist, "Syngeneic tobacco carcinogen-induced mouse lung adenocarcinoma model exhibits PD-L1 expression and high tumor mutational burden," *JCI Insight*, vol. 6, no. 3, Feb. 2021. 76

[144] F. W. Stearns, "One hundred years of pleiotropy: a retrospective," *Genetics*, vol. 186, no. 3, pp. 767–773, Nov. 2010. 4, 6

[145] M. K. Stein, M. Pandey, J. Xiu, H. Tae, J. Swensen, S. Mittal, A. J. Brenner, W. M. Korn, A. B. Heimberger, and M. G. Martin, "Tumor mutational burden is site specific in Non–Small-Cell lung cancer and is highest in lung adenocarcinoma brain metastases," *JCO Precision Oncology*, no. 3, pp. 1–13, Dec 2019. 77

[146] J. H. Strickler, B. A. Hanks, and M. Khasraw, "Tumor mutational burden as a predictor of immunotherapy response: Is more always better?" *Clin. Cancer Res.*, vol. 27, no. 5, pp. 1236–1241, Mar. 2021. 100

[147] C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray, B. Liu, P. Matthews, G. Ong, J. Pell, A. Silman, A. Young, T. Sprosen, T. Peakman, and R. Collins, "UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age," *PLoS Med.*, vol. 12, no. 3, p. e1001779, Mar. 2015. 3, 7, 17, 24, 31, 48, 51

[148] X. Sun, A. Xue, T. Qi, D. Chen, D. Shi, Y. Wu, Z. Zheng, J. Zeng, and J. Yang, "Tumor mutational burden is polygenic and genetically associated with complex traits and diseases," *Cancer Research*, vol. 81, no. 5, pp. 1230–1239, March 2021. 77, 89, 90

[149] E. Svensson, C. F. Christiansen, S. P. Ulrichsen, M. R. Rørth, and H. T. Sørensen, "Survival after bone metastasis by primary cancer type: a danish population-based cohort study," *BMJ Open*, vol. 7, no. 9, p. e016022, Sep. 2017. 94

[150] S. Turajlic, K. Litchfield, H. Xu, R. Rosenthal, N. McGranahan, J. L. Reading, Y. N. S. Wong, A. Rowan, N. Kanu, M. Al Bakir, T. Chambers, R. Salgado, P. Savas, S. Loi, N. J. Birkbak, L. Sansregret, M. Gore, J. Larkin, S. A. Quezada, and C. Swanton, "Insertion-and-deletion-derived tumour-specific neoantigens and the immunogenic phenotype: a pan-cancer analysis," *Lancet Oncol.*, vol. 18, no. 8, pp. 1009–1021, Aug. 2017. 76

[151] P. Turley, R. K. Walters, O. Maghzian, A. Okbay, J. J. Lee, M. A. Fontana, T. A. Nguyen-Viet, R. Wedow, M. Zacher, N. A. Furlotte, 23andMe Research Team, Social Science Genetic Association Consortium, P. Magnusson, S. Oskarsson, M. Johannesson, P. M. Visscher, D. Laibson, D. Cesarini, B. M. Neale, and D. J. Benjamin, "Multi-trait analysis of genome-wide association summary statistics using MTAG," *Nat. Genet.*, vol. 50, no. 2, pp. 229–237, Feb. 2018. 6, 7, 15, 24

[152] Y. Van Herck, A. Feyaerts, S. Alibhai, D. Papamichael, L. Decoster, Y. Lambrechts, M. Pinchuk, O. Bechter, J. Herrera-Caceres, F. Bibeau, C. Desmedt, S. Hatse, and H. Wildiers, "Is cancer biology different in older patients?" *The Lancet Healthy Longevity*, vol. 2, no. 10, pp. e663–e677, Oct. 2021. 92, 94

[153] W. van Rheenen, W. Peyrot, A. J. Schork, S. Hong Lee, and W. N. R, "Genetic correlations of polygenic disease traits: from theory to practice," *Nat. Rev. Genet*, vol. 20, pp. 567–581, 2019. 8

[154] D. S. Vinay, E. P. Ryan, G. Pawelec, W. H. Talib, J. Stagg, E. Elkord, T. Lichtor, W. K. Decker, R. L. Whelan, H. M. C. S. Kumara, E. Signori, K. Honoki, A. G. Georgakilas, A. Amin, W. G. Helferich, C. S. Boosani, G. Guha, M. R. Ciriolo, S. Chen, S. I. Mohammed, A. S. Azmi, W. N. Keith, A. Bilsland, D. Bhakta, D. Halicka, H. Fujii, K. Aquilano, S. S. Ashraf, S. Nowsheen, X. Yang, B. K. Choi, and B. S. Kwon, "Immune evasion in cancer: Mechanistic basis and therapeutic strategies," *Semin. Cancer Biol.*, vol. 35, pp. S185–S198, Dec. 2015. 95

[155] P. M. Visscher, "A note on the asymptotic distribution of likelihood ratio tests to test variance components," *Twin Res. Hum. Genet.*, vol. 9, no. 4, pp. 490–495, Aug. 2006. 8

[156] P. M. Visscher and J. Yang, "A plethora of pleiotropy across complex traits," *Nat. Genet.*, vol. 48, no. 7, pp. 707–708, Jun. 2016. 4, 6

[157] P. M. Visscher, N. R. Wray, Q. Zhang, P. Sklar, M. I. McCarthy, M. A. Brown, and J. Yang, "10 years of gwas discovery: biology, function, and translation," *The American Journal of Human Genetics*, vol. 101, no. 1, pp. 5–22, 2017. 2, 3, 4

[158] G. Wang, A. Sarkar, P. Carbonetto, and M. Stephens, "A simple new approach to variable selection in regression, with application to genetic fine-mapping," *bioRxiv*, p. 501114, Jun. 2020. 54, 62, 112

[159] X. Wang, B. Ricciuti, T. Nguyen, X. Li, M. S. Rabin, M. M. Awad, X. Lin, B. E. Johnson, and D. C. Christiani, "Association between smoking history and tumor mutation burden in advanced Non-Small cell lung cancer," *Cancer Res.*, vol. 81, no. 9, pp. 2566–2573, May 2021. 76

[160] Z. Wang, B.-Y. Liao, and J. Zhang, "Genomic patterns of pleiotropy and the evolution of complexity," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 107, no. 42, pp. 18 034–18 039, Oct. 2010. 4, 6

[161] I. R. Watson, K. Takahashi, P. A. Futreal, and L. Chin, "Emerging patterns of somatic mutations in cancer," *Nat. Rev. Genet.*, vol. 14, no. 10, pp. 703–718, Oct. 2013. 76

[162] J. L. Weissfeld, Y. Lin, H.-M. Lin, B. F. Kurland, D. O. Wilson, C. R. Fuhrman, A. Pennathur, M. Romkes, T. Nukui, J.-M. Yuan, J. M. Siegfried, and B. Diergaarde, "Lung cancer risk prediction using common SNPs located in GWAS-Identified susceptibility regions," *J. Thorac. Oncol.*, vol. 10, no. 11, pp. 1538–1545, Nov. 2015. 76

[163] G. L. Wojcik, M. Graff, K. K. Nishimura, R. Tao, J. Haessler, C. R. Gignoux, H. M. Highland, Y. M. Patel, E. P. Sorokin, C. L. Avery, G. M. Belbin, S. A. Bien, I. Cheng, S. Cullina, C. J. Hodonsky, Y. Hu, L. M. Huckins, J. Jeff, A. E. Justice, J. M. Kocarnik, U. Lim, B. M. Lin, Y. Lu, S. C. Nelson, S.-S. L. Park, H. Poisner, M. H. Preuss, M. A. Richard, C. Schurmann, V. W. Setiawan, A. Sockell, K. Vahi, M. Verbanck, A. Vishnu, R. W. Walker, K. L. Young, N. Zubair, V. Acuña-Alonso, J. L. Ambite, K. C. Barnes, E. Boerwinkle, E. P. Bottinger, C. D. Bustamante, C. Caberto, S. Canizales-Quinteros, M. P. Conomos, E. Deelman, R. Do, K. Doheny, L. Fernández-Rhodes, M. Fornage, B. Hailu, G. Heiss, B. M. Henn, L. A. Hindorff, R. D. Jackson, C. A. Laurie, C. C. Laurie, Y. Li, D.-Y. Lin, A. Moreno-Estrada, G. Nadkarni, P. J. Norman, L. C. Pooler, A. P. Reiner, J. Romm, C. Sabatti, K. Sandoval, X. Sheng, E. A. Stahl, D. O. Stram, T. A. Thornton, C. L. Wassel, L. R. Wilkens, C. A. Winkler, S. Yoneyama, S. Buyske, C. A. Haiman, C. Kooperberg, L. Le Marchand, R. J. F. Loos, T. C. Matise, K. E. North, U. Peters, E. E. Kenny, and C. S. Carlson, "Genetic analyses of diverse popula-

tions improves discovery for complex traits," *Nature*, vol. 570, no. 7762, pp. 514–518, Jun. 2019. 46, 67

[164] S. L. Wong, N. Gu, M. Banerjee, J. D. Birkmeyer, and N. J. Birkmeyer, "The impact of socioeconomic status on cancer care and survival," *J. Clin. Orthod.*, vol. 29, no. 15_suppl, pp. 6004–6004, May 2011. 90

[165] A. R. Wood, T. Esko, J. Yang, S. Vedantam, T. H. Pers, S. Gustafsson, A. Y. Chu, K. Estrada, J. Luan, Z. Kutalik, N. Amin, M. L. Buchkovich, D. C. Croteau-Chonka, F. R. Day, Y. Duan, T. Fall, R. Fehrmann, T. Ferreira, A. U. Jackson, J. Karjalainen, K. S. Lo, A. E. Locke, R. Mägi, E. Mihailov, E. Porcu, J. C. Randall, A. Scherag, A. A. E. Vinkhuyzen, H.-J. Westra, T. W. Winkler, T. Workalemahu, J. H. Zhao, D. Absher, E. Albrecht, D. Anderson, J. Baron, M. Beekman, A. Demirkan, G. B. Ehret, B. Feenstra, M. F. Feitosa, K. Fischer, R. M. Fraser, A. Goel, J. Gong, A. E. Justice, S. Kanoni, M. E. Kleber, K. Kristiansson, U. Lim, V. Lotay, J. C. Lui, M. Mangino, I. Mateo Leach, C. Medina-Gomez, M. A. Nalls, D. R. Nyholt, C. D. Palmer, D. Pasko, S. Pechlivanis, I. Prokopenko, J. S. Ried, S. Ripke, D. Shungin, A. Stancáková, R. J. Strawbridge, Y. J. Sung, T. Tanaka, A. Teumer, S. Trompet, S. W. van der Laan, J. van Setten, J. V. Van Vliet-Ostaptchouk, Z. Wang, L. Yengo, W. Zhang, U. Afzal, J. Arnlöv, G. M. Arscott, S. Bandinelli, A. Barrett, C. Bellis, A. J. Bennett, C. Berne, M. Blüher, J. L. Bolton, Y. Böttcher, H. A. Boyd, M. Bruinenberg, B. M. Buckley, S. Buyske, I. H. Caspersen, P. S. Chines, R. Clarke, S. Claudi-Boehm, M. Cooper, E. W. Daw, P. A. De Jong, J. Deelen, G. Delgado, J. C. Denny, R. Dhonukshe-Rutten, M. Dimitriou, A. S. F. Doney, M. Dörr, N. Eklund, E. Eury, L. Folkersen, M. E. Garcia, F. Geller, V. Giedraitis, A. S. Go, H. Grallert, T. B. Grammer, J. Gräßler, H. Grönberg, L. C. P. G. M. de Groot, C. J. Groves, J. Haessler, P. Hall, T. Haller, G. Hallmans, A. Hannemann, C. A. Hartman, M. Hassinen, C. Hayward, N. L. Heard-Costa, Q. Helmer, G. Hemani, A. K. Henders, H. L. Hillege, M. A. Hlatky, W. Hoffmann, P. Hoffmann, O. Holmen, J. J. Houwing-Duistermaat, T. Illig, A. Isaacs, A. L. James, J. Jeff, B. Johansen, Å. Johansson, J. Jolley, T. Juliusdottir,

J. Junttila, A. N. Kho, L. Kinnunen, N. Klopp, T. Kocher, W. Kratzer, P. Lichtner, L. Lind, J. Lindström, S. Lobbens, M. Lorentzon, Y. Lu, V. Lyssenko, P. K. E. Magnusson, A. Mahajan, M. Maillard, W. L. McArdle, C. A. McKenzie, S. McLachlan, P. J. McLaren, C. Menni, S. Merger, L. Milani, A. Moayyeri, K. L. Monda, M. A. Morken, G. Müller, M. Müller-Nurasyid, A. W. Musk, N. Narisu, M. Nauck, I. M. Nolte, M. M. Nöthen, L. Oozageer, S. Pilz, N. W. Rayner, F. Renstrom, N. R. Robertson, L. M. Rose, R. Roussel, S. Sanna, H. Scharnagl, S. Scholtens, F. R. Schumacher, H. Schunkert, R. A. Scott, J. Sehmi, T. Seufferlein, J. Shi, K. Silventoinen, J. H. Smit, A. V. Smith, J. Smolonska, A. V. Stanton, K. Stirrups, D. J. Stott, H. M. Stringham, J. Sundström, M. A. Swertz, A.-C. Syvänen, B. O. Tayo, G. Thorleifsson, J. P. Tyrer, S. van Dijk, N. M. van Schoor, N. van der Velde, D. van Heemst, F. V. A. van Oort, S. H. Vermeulen, N. Verweij, J. M. Vonk, L. L. Waite, M. Waldenberger, R. Wennauer, L. R. Wilkens, C. Willenborg, T. Wilsgaard, M. K. Wojczynski, A. Wong, A. F. Wright, Q. Zhang, D. Arveiler, S. J. L. Bakker, J. Beilby, R. N. Bergman, S. Bergmann, R. Biffar, J. Blangero, D. I. Boomsma, S. R. Bornstein, P. Bovet, P. Brambilla, M. J. Brown, H. Campbell, M. J. Caulfield, A. Chakravarti, R. Collins, F. S. Collins, D. C. Crawford, L. A. Cupples, J. Danesh, U. de Faire, H. M. den Ruijter, R. Erbel, J. Erdmann, J. G. Eriksson, M. Farrall, E. Ferrannini, J. Ferrières, I. Ford, N. G. Forouhi, T. Forrester, R. T. Gansevoort, P. V. Gejman, C. Gieger, A. Golay, O. Gottesman, V. Gudnason, U. Gyllensten, D. W. Haas, A. S. Hall, T. B. Harris, A. T. Hattersley, A. C. Heath, C. Hengstenberg, A. A. Hicks, L. A. Hindorff, A. D. Hingorani, A. Hofman, G. K. Hovingh, S. E. Humphries, S. C. Hunt, E. Hypponen, K. B. Jacobs, M.-R. Jarvelin, P. Jousilahti, A. M. Jula, J. Kaprio, J. J. P. Kastelein, M. Kayser, F. Kee, S. M. Keinanen-Kiukaanniemi, L. A. Kiemeney, J. S. Kooner, C. Kooperberg, S. Koskinen, P. Kovacs, A. T. Kraja, M. Kumari, J. Kuusisto, T. A. Lakka, C. Langenberg, L. Le Marchand, T. Lehtimäki, S. Lupoli, P. A. F. Madden, S. Männistö, P. Manunta, A. Marette, T. C. Matise, B. McKnight, T. Meitinger, F. L. Moll, G. W. Montgomery, A. D. Morris, A. P. Morris, J. C. Murray, M. Nelis, C. Ohlsson, A. J. Oldehinkel, K. K. Ong, W. H. Ouwehand, G. Pasterkamp, A. Peters, P. P. Pramstaller,

J. F. Price, L. Qi, O. T. Raitakari, T. Rankinen, D. C. Rao, T. K. Rice, M. Ritchie, I. Rudan, V. Salomaa, N. J. Samani, J. Saramies, M. A. Sarzynski, P. E. H. Schwarz, S. Sebert, P. Sever, A. R. Shuldiner, J. Sinisalo, V. Steinthorsdottir, R. P. Stolk, J.-C. Tardif, A. Tönjes, A. Tremblay, E. Tremoli, J. Virtamo, M.-C. Vohl, Electronic Medical Records and Genomics (eMEMERGEGE) Consortium, MIGen Consortium, PAGEGE Consortium, LifeLines Cohort Study, P. Amouyel, F. W. Asselbergs, T. L. Assimes, M. Bochud, B. O. Boehm, E. Boerwinkle, E. P. Bottinger, C. Bouchard, S. Cauchi, J. C. Chambers, S. J. Chanock, R. S. Cooper, P. I. W. de Bakker, G. Dedoussis, L. Ferrucci, P. W. Franks, P. Froguel, L. C. Groop, C. A. Haiman, A. Hamsten, M. G. Hayes, J. Hui, D. J. Hunter, K. Hveem, J. W. Jukema, R. C. Kaplan, M. Kivimaki, D. Kuh, M. Laakso, Y. Liu, N. G. Martin, W. März, M. Melbye, S. Moebus, P. B. Munroe, I. Njølstad, B. A. Oostra, C. N. A. Palmer, N. L. Pedersen, M. Perola, L. Pérusse, U. Peters, J. E. Powell, C. Power, T. Quertermous, R. Rauramaa, E. Reinmaa, P. M. Ridker, F. Rivadeneira, J. I. Rotter, T. E. Saaristo, D. Saleheen, D. Schlessinger, P. E. Slagboom, H. Snieder, T. D. Spector, K. Strauch, M. Stumvoll, J. Tuomilehto, M. Uusitupa, P. van der Harst, H. Völzke, M. Walker, N. J. Wareham, H. Watkins, H.-E. Wichmann, J. F. Wilson, P. Zanen, P. Deloukas, I. M. Heid, C. M. Lindgren, K. L. Mohlke, E. K. Speliotes, U. Thorsteinsdottir, I. Barroso, C. S. Fox, K. E. North, D. P. Strachan, J. S. Beckmann, S. I. Berndt, M. Boehnke, I. B. Borecki, M. I. McCarthy, A. Metspalu, K. Stefansson, A. G. Uitterlinden, C. M. van Duijn, L. Franke, C. J. Willer, A. L. Price, G. Lettre, R. J. F. Loos, M. N. Weedon, E. Ingelsson, J. R. O'Connell, G. R. Abecasis, D. I. Chasman, M. E. Goddard, P. M. Visscher, J. N. Hirschhorn, and T. M. Frayling, "Defining the role of common variation in the genomic and biological architecture of adult human height," *Nat. Genet.*, vol. 46, no. 11, pp. 1173–1186, Nov. 2014. 27

[166] N. R. Wray, K. E. Kemper, B. J. Hayes, M. E. Goddard, and P. M. Visscher, "Complex trait prediction from genome data: Contrasting EBV in livestock to PRS in humans: Genomic prediction," *Genetics*, vol. 211, no. 4, pp. 1131–1141, Apr. 2019. 89

[167] C. B. Xavier, G. Guardia, C. D. H. Lopes, B. M. Awni, E. F. Campos, J. P. Alves, A. A. Camargo, P. A. F. Galante, and D. L. Jardim, "Association between tumor mutational burden (TMB) and mutational profile and its effect on overall survival: A post hoc analysis of patients with TMB-high and TMB-low metastatic cancer treated with immune checkpoint inhibitors (ICI)," *J. Clin. Orthod.*, vol. 40, no. 16_suppl, pp. 2632–2632, Jun. 2022. 92

[168] J. Yang, S. H. Lee, M. E. Goddard, and P. M. Visscher, "GCTA: a tool for genome-wide complex trait analysis," *Am. J. Hum. Genet.*, vol. 88, no. 1, pp. 76–82, Jan. 2011. 49

[169] C. X. Yap, J. Sidorenko, Y. Wu, K. E. Kemper, J. Yang, N. R. Wray, M. R. Robinson, and P. M. Visscher, "Dissection of genetic variation and evidence for pleiotropy in male pattern baldness," *Nat. Commun.*, vol. 9, no. 1, p. 5407, Dec. 2018. 6

[170] A. Zehir, R. Benayed, R. H. Shah, A. Syed, S. Middha, H. R. Kim, P. Srinivasan, J. Gao, D. Chakravarty, S. M. Devlin, M. D. Hellmann, D. A. Barron, A. M. Schram, M. Hameed, S. Dogan, D. S. Ross, J. F. Hechtman, D. F. DeLair, J. Yao, D. L. Mandelker, D. T. Cheng, R. Chandramohan, A. S. Mohanty, R. N. Ptashkin, G. Jayakumaran, M. Prasad, M. H. Syed, A. B. Rema, Z. Y. Liu, K. Nafa, L. Borsu, J. Sadowska, J. Casanova, R. Bacares, I. J. Kiecka, A. Razumova, J. B. Son, L. Stewart, T. Baldi, K. A. Mullaney, H. Al-Ahmadie, E. Vakiani, A. A. Abeshouse, A. V. Penson, P. Jonsson, N. Camacho, M. T. Chang, H. H. Won, B. E. Gross, R. Kundra, Z. J. Heins, H.-W. Chen, S. Phillips, H. Zhang, J. Wang, A. Ochoa, J. Wills, M. Eubank, S. B. Thomas, S. M. Gardos, D. N. Reales, J. Galle, R. Durany, R. Cambria, W. Abida, A. Cercek, D. R. Feldman, M. M. Gounder, A. A. Hakimi, J. J. Harding, G. Iyer, Y. Y. Janjigian, E. J. Jordan, C. M. Kelly, M. A. Lowery, L. G. T. Morris, A. M. Omuro, N. Raj, P. Razavi, A. N. Shoushtari, N. Shukla, T. E. Soumerai, A. M. Varghese, R. Yaeger, J. Coleman, B. Bochner, G. J. Riely, L. B. Saltz, H. I. Scher, P. J. Sabbatini, M. E. Robson, D. S. Klimstra, B. S. Taylor, J. Baselga, N. Schultz, D. M. Hyman, M. E. Arcila, D. B. Solit, M. Ladanyi, and M. F. Berger, "Mutational landscape of metastatic

cancer revealed from prospective clinical sequencing of 10,000 patients," *Nat. Med.*, vol. 23, no. 6, pp. 703–713, Jun. 2017. 76

[171] H. Zhang, T. U. Ahearn, J. Lecarpentier, D. Barnes, J. Beesley, G. Qi, X. Jiang, T. A. O'Mara, N. Zhao, M. K. Bolla, A. M. Dunning, J. Dennis, Q. Wang, Z. A. Ful, K. Aittomäki, I. L. Andrulis, H. Anton-Culver, V. Arndt, K. J. Aronson, B. K. Arun, P. L. Auer, J. Azzollini, D. Barrowdale, H. Becher, M. W. Beckmann, S. Behrens, J. Benitez, M. Bermisheva, K. Bialkowska, A. Blanco, C. Blomqvist, N. V. Bogdanova, S. E. Bojesen, B. Bonanni, D. Bondavalli, A. Borg, H. Brauch, H. Brenner, I. Briceno, A. Broeks, S. Y. Brucker, T. Brüning, B. Burwinkel, S. S. Buys, H. Byers, T. Caldés, M. A. Caligo, M. Calvello, D. Campa, J. E. Castelao, J. Chang-Claude, S. J. Chanock, M. Christiaens, H. Christiansen, W. K. Chung, K. B. M. Claes, C. L. Clarke, S. Cornelissen, F. J. Couch, A. Cox, S. S. Cross, K. Czene, M. B. Daly, P. Devilee, O. Diez, S. M. Domchek, T. Dörk, M. Dwek, D. M. Eccles, A. B. Ekici, D. G. Evans, P. A. Fasching, J. Figueroa, L. Foretova, F. Fostira, E. Friedman, D. Frost, M. Gago-Dominguez, S. M. Gapstur, J. Garber, J. A. García-Sáenz, M. M. Gaudet, S. A. Gayther, G. G. Giles, A. K. Godwin, M. S. Goldberg, D. E. Goldgar, A. González-Neira, M. H. Greene, J. Gronwald, P. Guénel, L. Häberle, E. Hahnen, C. A. Haiman, C. R. Hake, P. Hall, U. Hamann, E. F. Harkness, B. A. M. Heemskerk-Gerritsen, P. Hillemanns, F. B. L. Hogervorst, B. Holleczek, A. Hollestelle, M. J. Hooning, R. N. Hoover, J. L. Hopper, A. Howell, H. Huebner, P. J. Hulick, E. N. Imyanitov, kConFab Investigators, ABCTB Investigators, C. Isaacs, L. Izatt, A. Jager, M. Jakimovska, A. Jakubowska, P. James, R. Janavicius, W. Janni, E. M. John, M. E. Jones, A. Jung, R. Kaaks, P. M. Kapoor, B. Y. Karlan, R. Keeman, S. Khan, E. Khusnutdinova, C. M. Kitahara, Y.-D. Ko, I. Konstantopoulou, L. B. Koppert, S. Koutros, V. N. Kristensen, A.-V. Laenkholm, D. Lambrechts, S. C. Larsson, P. Laurent-Puig, C. Lazaro, E. Lazarova, F. Lejbkowicz, G. Leslie, F. Lesueur, A. Lindblom, J. Lissowska, W.-Y. Lo, J. T. Loud, J. Lubinski, A. Lukomska, R. J. MacInnis, A. Mannermaa, M. Manoochehri, S. Manoukian, S. Margolin, M. E. Martinez, L. Matricardi, L. McGuffog, C. McLean,

N. Mebirouk, A. Meindl, U. Menon, A. Miller, E. Mingazheva, M. Montagna, A. M. Mulligan, C. Mulot, T. A. Muranen, K. L. Nathanson, S. L. Neuhausen, H. Nevanlinna, P. Neven, W. G. Newman, F. C. Nielsen, L. Nikitina-Zake, J. Nodora, K. Offit, E. Olah, O. I. Olopade, H. Olsson, N. Orr, L. Papi, J. Papp, T.-W. Park-Simon, M. T. Parsons, B. Peissel, A. Peixoto, B. Peshkin, P. Peterlongo, J. Peto, K.-A. Phillips, M. Piedmonte, D. Plaseska-Karanfilska, K. Prajzendanc, R. Prentice, D. Prokofyeva, B. Rack, P. Radice, S. J. Ramus, J. Rantala, M. U. Rashid, G. Rennert, H. S. Rennert, H. A. Risch, A. Romero, M. A. Rookus, M. Rübner, T. Rüdiger, E. Saloustros, S. Sampson, D. P. Sandler, E. J. Sawyer, M. T. Scheuner, R. K. Schmutzler, A. Schneeweiss, M. J. Schoemaker, B. Schöttker, P. Schürmann, L. Senter, P. Sharma, M. E. Sherman, X.-O. Shu, C. F. Singer, S. Smichkoska, P. Soucy, M. C. Southey, J. J. Spinelli, J. Stone, D. Stoppa-Lyonnet, EMBRACE Study, GEMO Study Collaborators, A. J. Swerdlow, C. I. Szabo, R. M. Tamimi, W. J. Tapper, J. A. Taylor, M. R. Teixeira, M. Terry, M. Thomassen, D. L. Thull, M. Tischkowitz, A. E. Toland, R. A. E. M. Tollenaar, I. Tomlinson, D. Torres, M. A. Troester, T. Truong, N. Tung, M. Untch, C. M. Vachon, A. M. W. van den Ouweland, L. E. van der Kolk, E. M. van Veen, E. J. van-Rensburg, A. Vega, B. Wappenschmidt, C. R. Weinberg, J. N. Weitzel, H. Wildiers, R. Winqvist, A. Wolk, X. R. Yang, D. Yannoukakos, W. Zheng, K. K. Zorn, R. L. Milne, P. Kraft, J. Simard, P. D. P. Pharoah, K. Michailidou, A. C. Antoniou, M. K. Schmidt, G. Chenevix-Trench, D. F. Easton, N. Chatterjee, and M. García-Closas, "Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses," *Nat. Genet.*, vol. 52, no. 6, pp. 572–581, Jun. 2020. 89, 105

[172] W. Zhang, X. Gao, X. Shi, B. Zhu, Z. Wang, H. Gao, L. Xu, L. Zhang, J. Li, and Y. Chen, "PCA-Based Multiple-Trait GWAS analysis: A powerful model for exploring pleiotropy," *Animals (Basel)*, vol. 8, no. 12, Dec. 2018. 6

[173] Y. Zhao, X. Fu, J. I. Lopez, A. Rowan, L. Au, A. Fendler, S. Hazell, H. Xu, S. Horswell, S. T. C. Shepherd, L. Spain, F. Byrne, G. Stamp, T. O'Brien, D. Nicol, M. Augustine,

A. Chandra, S. Rudman, A. Toncheva, L. Pickering, E. Sahai, J. Larkin, P. A. Bates, C. Swanton, S. Turajlic, TRACERx Renal Consortium, and K. Litchfield, "Selection of metastasis competent subclones in the tumour interior," *Nat Ecol Evol*, vol. 5, no. 7, pp. 1033–1045, Jul. 2021. 77

[174] X. Zhou and M. Stephens, "Efficient multivariate linear mixed model algorithms for genome-wide association studies," *Nat. Methods*, vol. 11, no. 4, pp. 407–409, Apr. 2014. 6

[175] X. Zhu, T. Feng, B. O. Tayo, J. Liang, J. H. Young, N. Franceschini, J. A. Smith, L. R. Yanek, Y. V. Sun, T. L. Edwards, W. Chen, M. Nalls, E. Fox, M. Sale, E. Bottinger, C. Rotimi, COGENT BP Consortium, Y. Liu, B. McKnight, K. Liu, D. K. Arnett, A. Chakravati, R. S. Cooper, and S. Redline, "Meta-analysis of correlated traits via summary statistics from GWASs with an application in hypertension," *Am. J. Hum. Genet.*, vol. 96, no. 1, pp. 21–36, Jan. 2015. 6