

Genetic Prediction of Complex Traits Across Diverse Populations

by
Taylor Cavazos

DISSERTATION
Submitted in partial satisfaction of the requirements for degree of
DOCTOR OF PHILOSOPHY

in

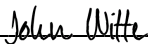
Biological and Medical Informatics

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

DocuSigned by:

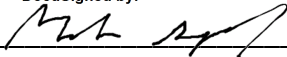


5991B9E9C971446...

John Witte

Chair

DocuSigned by:



DocuSigned by:



D11C633BFE5F4F9...

Mark Segal

Elad Ziv

Committee Members

Copyright 2022

by

Taylor Cavazos

DEDICATION

To my mom, who always taught me to follow my dreams. To my dad, who showed me that with hard work you can accomplish anything. To my sister, who is a model of compassion and a reliable source of laughter. I am blessed to have a truly supportive family who have been with me every step of the way.

ACKNOWLEDGEMENTS

My mom recently told me, “You aren’t one to choose the easy path.” The choice to pursue a PhD came with many challenges, but also immense personal growth and incredible opportunities that have prepared me for a lifelong career in Bioinformatics. However, it is not lost on me that choosing the path of greater resistance has only been possible thanks to the many mentors, family, and friends who have supported me. I am eternally grateful for each and every one of you.

First, I want to thank my PhD advisor, **John Witte**. You have created a research environment that motivates curiosity and creativity while being supportive and caring. Before joining your group I had never been asked what I wanted out of graduate school. You not only gave me the space to figure this out, but also the tools to prioritize and accomplish my research goals. Without your encouragement to develop my own research projects, I may not have had the chance to explore areas I was truly passionate about or gain the skills necessary to become a more independent scientist.

Secondly, I want to thank my thesis committee members **Elad Ziv** and **Mark Segal**. Elad, I greatly appreciated your new ideas, clinical insights, and candor. You inspired me to think outside of the box and helped me realize the impact of my work. Mark, you are an incredible statistician who made sure I conducted rigorous science. I appreciate that you kept my career interests in mind and I’m grateful you connected me to an internship opportunity at 23andMe, which greatly enriched my PhD experience.

And to the friends I’ve gained through the Witte Lab, you have all been tremendously helpful over the years. **Nima Emami** and **Clint Cario**, thank you both for your immense computational support and for being great examples of what a successful PhD looks like in our program. **Emmalyn Chen** and **Linda Kachuri**, you’ve both served as the best role models and sources of advice throughout my PhD, but more than that have become incredible lifelong friends. Thank you both for being someone I could turn to whenever things were difficult and for being a part of my home away from home.

Of course, I can't forget my mentors at UCSD who inspired me to pursue a graduate degree in Bioinformatics in the first place. First, **Pablo Tamayo** and **Jill Mesirov**, you both had a tremendous impact on my early scientific career. As a first-generation college student, the process of pursuing higher education was completely lost on me. You both provided opportunities that gave me confidence and helped mold me into the scientist I am today. Also, **Niema Moshiri** who spent countless hours helping me work through coding exercises. I am incredibly grateful for your support and have enjoyed watching you transition into a professor; your current (and future) students are very lucky to have you as a mentor.

Owen Cockerill, you have been with me through it all; from being a welcome distraction when I had to prepare for graduate school applications, to supporting and encouraging my decision to come to UCSF and dealing with my many ups and downs throughout the PhD process. You are my rock and my constant reminder to live in the present, enjoy the little moments, and not take life too seriously. **Kathy and Craig Cockerill**, thank you for welcoming me into your family, introducing me to new experiences and places, and being there to celebrate my accomplishments, no matter how big or small they were.

And finally, my immediate and extended family who excitedly supported me throughout my PhD, while being a constant reminder of where I came from. To my late grandpa **Richard Taliaferro**, you were always so proud of me and made me feel like I could accomplish anything. And to my nana and grandpa, **Ophelia** and **Manuel Cavazos**, I am extremely grateful for the countless hours I spent with you growing up and for the foundation you both laid so that your children and grandchildren could be successful. And of course, my parents and sister, who I dedicate this work to. You gave me the courage to follow my dreams and helped me in every way to make them a reality. You are truly the best support system and have put up with far more than you deserve to help get me to where I am today. Without you all, none of this would have been possible.

CONTRIBUTIONS

Chapter 2 of this thesis is a reprint of materials as it appears in the journal, *Human Genetics and Genomics Advances* [1]. Chapter 3 of this thesis is a reprint of materials as it appears on *medRxiv* [2]. Co-authors listed in these works provided feedback and discussion and the corresponding author, John S. Witte, supervised each of the works listed that form the basis for this thesis. The published material is substantially the byproduct of Taylor B. Cavazos' PhD research at the University of California, San Francisco and was largely developed and conducted by her. Each of the published works are equivalent to a standard thesis chapter.

[1] Cavazos TB, Witte JS. Inclusion of variants discovered from diverse populations improves polygenic risk score transferability. *HGG Adv.* 2021;2(1):100017. doi:10.1016/j.xhgg.2020.100017

[2] Cavazos TB, Kachuri L, Graff RE, Nierenberg JL, *et al.* Assessment of Genetic Susceptibility to Multiple Primary Cancers through Whole-Exome Sequencing in Two Large Multi-Ancestry Studies. *medRxiv.* 2022; 02.11.22270688. doi:10.1101/2022.02.11.22270688

The only true wisdom

is in knowing

you know nothing

– Socrates

ABSTRACT

Genetic Prediction of Complex Traits Across Diverse Populations

Taylor Cavazos

Background: Discoveries made through genome-wide association studies have revolutionized the field of human genetics by uncovering disease mechanisms and enabling precision medicine. Although there has been tremendous effort to obtain cohort sizes on the order of hundreds of thousands of individuals, there is an immense underrepresentation of non-European ancestries, which has potential to contribute to health inequity. Here, we present works on (i) an in-depth simulation to identify optimal approaches for achieving equitable accuracy for polygenic risk scores (PRS) in diverse populations and (ii) an application of trans-ancestry discovery of germline associations in the complex phenotype of multiple primary tumors.

Methods: (i) Through our simulation framework, we implement many strategies for building PRS, including a local-ancestry specific approach, and measure accuracy in admixed and African ancestry individuals. (ii) We also conducted a whole-exome sequencing study of two large, multi-ancestry populations consisting of 6,429 multiple cancer cases, 29,091 single cancer cases, and 165,853 cancer-free controls. We employed single-variant and gene-based tests to characterize the genetic susceptibility to multiple primary tumors in comparison to individuals with one and, separately, no cancers through the investigation of rare and common variation. **Results and Conclusions:** (i) Variants discovered in African ancestry populations have greater potential to achieve unbiased PRS prediction across populations. Studies should prioritize the inclusion of diverse participants in GWAS, and care must be taken with the interpretation of currently available risk scores. (ii) Our applied trans-ancestry analysis of multiple primary tumors identifies rare loss-of-function variants and gene-level associations with cross-cancer pleiotropy and potential for prioritizing cancer survivors at high risk for developing subsequent tumors.

Keywords: polygenic risk scores, population genetics, statistical genetics, local ancestry, GWAS, whole-exome sequencing, cancer, pleiotropy

TABLE OF CONTENTS

CHAPTER I.....	1
Introduction.....	1
1.1 Genetic variation in humans.....	1
1.2 Improving disease prediction with germline genetics.....	3
1.3 Cancer genetics and pleiotropy	4
1.4 The lack of diversity in human genetic studies.....	5
References.....	8
CHAPTER II	11
Addressing the limitations of polygenic risk scores	11
2.1 Abstract	11
2.2 Introduction	11
2.3 Materials and Methods.....	13
2.3.1 Simulation of population genotypes.....	13
2.3.2 True and GWAS estimated polygenic risk scores.....	13
2.3.3 Multi-ancestry polygenic risk scores.....	15
2.3.4 Application to real data.....	16
2.4 Results	17
2.4.1 Generalizability of European-derived risk scores to an admixed population.....	17
2.4.2 Population-specific weighting of European selected variants.....	19
2.4.3 Performance of non-European PRS variant selection and weighting approaches.....	20
2.4.4 Inclusion of variants from diverse populations	21
2.4.5 Allele frequency and LD of GWAS variants.....	22
2.4.6 Application to real data.....	24
2.5 Discussion	24
2.6 Acknowledgements	28
2.7 Tables	30
2.8 Figures.....	31
2.9 Supplementary Materials.....	35
References.....	39
CHAPTER III	42
Genetic susceptibility to multiple primary tumors	42
3.1 Abstract	42
3.2 Introduction	42
3.3 Materials and Methods.....	43
3.3.1 Study populations and phenotyping	43

3.3.2	Genetic ancestry and principal components analysis	44
3.3.3	Whole-exome sequencing and quality control	45
3.3.4	Association analysis in individuals with multiple cancers versus cancer-free controls	46
3.3.5	Distinguishing susceptibility signals for multiple cancers versus single cancers	47
3.4	Results	48
3.4.1.	Characterization of multiple primary cancer diagnoses in two large study populations...	48
3.4.2.	Exome-wide single variant association analyses.....	48
3.4.3.	Gene-based analyses of multiple cancers	50
3.4.4.	Comparison of mutation burden in individuals with multiple versus single cancers.....	51
3.5	Discussion	52
3.6	Acknowledgements	55
3.7	Tables	56
3.8	Figures.....	57
3.9	Supplementary Materials.....	60
	References.....	72
	Funding and Support.....	76

LIST OF FIGURES

Figure 2.1 Accuracy of European derived PRSs by proportion of total ancestry.....	31
Figure 2.2 PRS construction approaches and performance in admixed individuals	32
Figure 2.3 Impact of African ancestry sample size on PRS accuracy and generalization.....	33
Figure 2.4 Allele frequency distribution of GWAS selected variants and LD tagging of causal variants .	34
Figure S2.1 European derived risk score accuracy with varying simulation causal variants and assumed heritability	36
Figure S2.2 PRS accuracy with varying p-value thresholds.....	37
Figure S2.3 Normalized LD tagging of causal variants by GWAS selected variants	38
Figure S2.4 HbA1c PRS performance in the UK Biobank.....	38
Figure 3.1 Cancer diagnosis pairs present in the combined study populations	57
Figure 3.2 Germline single variant association results for multiple primary cancers combined or grouped by organ site.....	58
Figure 3.3 Germline gene based association results for multiple primary cancers combined or grouped by organ site.....	59
Figure S3.1 Genetic ancestry in the Kaiser Permanente Research Bank and UK Biobank.....	62
Figure S3.2 Number of primary cancer diagnoses and time intervals between cancer diagnoses for Kaiser Permanente Research Bank and UK Biobank individuals with multiple cancers.....	66
Figure S3.3 Most common cancer pairs present in Kaiser Permanente and UK Biobank cases with multiple cancers	67
Figure S3.4 Cancers represented in the Kaiser Permanente Research Bank and UK Biobank with sufficient sample size for exome-wide association analyses	68
Figure S3.5 Significant single variant association results due to CHIP	69
Figure S3.6 Allele balance for findings related to lymphoid and myeloid neoplasms	70
Figure S3.7 Significant gene-based association results due to CHIP	71

LIST OF TABLES

Table 2.1 Summary of PRS variants and causal tagging across simulations.....	30
Table S2.1 Summary of HbA1c variants included for each PRS	35
Table 3.1 Summary of study populations	56
Table S3.1 Cancer definitions.....	60
Table S3.2 Cancer pairs with at least 25 cases combined across study populations	61
Table S3.3 Groupings of multiple cancer cases by a shared index cancer	62
Table S3.4 Single-variant case-control meta-analysis associations for multiple cancer phenotypes	63
Table S3.5 Gene-based case-control meta-analysis associations for multiple cancer phenotypes.....	63
Table S3.6 Case-case analysis for significant single-variant associations.....	64
Table S3.7 Case-case analysis for significant gene-based associations.....	64

CHAPTER I

Introduction

Many common and rare human diseases are influenced, at least partially, by genetic variation or changes in an individual's DNA sequence. Advances in techniques for measuring genetic variation, such as population scale genotyping and sequencing strategies, has led to the identification of thousands of genetic variants that may contribute to the development of specific traits and diseases. Disentangling the genetic contribution to disease is especially advantageous for personalizing screening prevention and treatment strategies; however, most human genetics studies have been conducted in individuals of European ancestry and discoveries may not translate to the greater diversity of global populations. In this chapter, I will detail the motivations for and complexity of studying human genetics, as well as describe the methodological and technological advances in the field that have enabled this work. I will also highlight coauthored studies that utilize germline genetics for disease prediction, specifically in cancer. Finally, I will summarize the current state of germline genetics research in populations of non-European ancestry, discuss the implications of underrepresentation in genetics studies, and show relevant examples where inclusion of diverse individuals leads to additional discovery.

1.1 Genetic variation in humans

The human genome acts as the blueprint for the human body and contains extensive information about human evolution, development, and disease. The hereditary genetic material, passed down from parent to child, is encoded as deoxyribonucleic acid (DNA) and is over 3 billion base pairs long. The ability to decipher and unlock discoveries from our genetic information can be attributed to a 13-year, international initiative by the Human Genome Project to sequence the first human genome^{1,2}. The development of a large-scale sequencing strategy, called shotgun sequencing³, and methodological advances in genome assembly were crucial for reading and determining the correct order of nucleotides in our DNA. In 2001,

these advances culminated in the landmark achievement of creating a consensus human DNA sequence, constructed from multiple individuals, that formed the basis of the human reference genome used today^{1,2}.

Since 2001, over 100,000 human genomes representing at least 26 global populations have been sequenced through research efforts such as the 1000 Genomes Project⁴, the International HapMap Project,⁵ the Haplotype Reference Consortium (HRC)⁶, and most recently the Trans-Omics for Precision Medicine (TOPMed)⁷ program among others. Comparisons across individual's genomes have shown that we are remarkably similar in terms of our genetic makeup, sharing ~99.9% of our DNA; however, the ~0.1% where we differ contains relevant insights into what makes us unique. Thus, through comparisons with the human reference genome, we can identify variation in the DNA sequence at either a single position or multiple positions, called single nucleotide polymorphisms (SNPs) or indels respectively. An ongoing challenge has been to link genetic variation found throughout human genomes to functional consequences, as well as determine which genetic variants influence physical traits, biomarkers, and disease.

In order to gain insights from DNA and measure its influence on complex traits, we need genetic information for tens to hundreds of thousands of individuals; unfortunately, the cost of whole-genome sequencing has been a limiting factor. It cost ~1 billion dollars to sequence the first human genome and although this has decreased substantially to ~\$1,000 per genome, sequencing a large number of individuals for each phenotype of interest is still impractical. A more efficient and cost-effective approach is through genotyping arrays, which measure targeted sites in the genome that can be used to infer missing sites (imputation); genotyping with imputation has been shown to be highly accurate for common variation (occurring in at least 1% of the population)⁸. Along with advances in genotyping technologies, the establishment of large population-based prospective cohorts has been crucial for facilitating genetic discovery. One example is the UK Biobank (UKB), an open-access resource of ~500,000 participants recruited between ages 40 to 69 with detailed phenotyping and longitudinal followup⁹. Another cohort, used in this dissertation and previously published work, is the Kaiser Permanente Research Bank (KPRB) which

includes members of Kaiser Permanente Northern California with extensive health information from self-reported surveys and electronic health records¹⁰. We can leverage these large-scale cohorts to conduct genome-wide association studies (GWAS), where regression models are used to test each position across the genome for an association with a trait of interest in a set of individuals¹¹.

1.2 Improving disease prediction with germline genetics

The vast amount of genetic and phenotypic data available has enabled the discovery of genetic variation with implications for both predicting and understanding mechanisms for rare and common disease. Early genetic studies focused on rare, monogenic/Mendelian diseases having a single disease-causing gene. This led to the discovery of *HBB* for sickle cell anemia, *CFTR* for cystic fibrosis, and *HTT* for Huntington's disease. Variation in these genes tend to be high penetrance, meaning a large proportion of individuals who are carriers for the disease variant will show the disease phenotype¹². In contrast, most common diseases tend to have a complex, polygenic architecture with hundreds to thousands of genetic variants across many genes all having smaller effects on the phenotype. Examples of diseases with a complex genetic architecture includes cancer, cardiovascular disease, and diabetes. While there are distinctions between Mendelian and complex disorders, these are often simplifications that do not fully convey the genetic and phenotypic variability of each trait^{13,14}. For example, breast cancer risk can be substantially increased due to a single high penetrance variant in *BRCA1/2*, or from many variants with small effects, or a combination of both¹⁵.

While genetic studies of Mendelian disorders are more likely to identify the true disease-causing variant, due to their high penetrance, it is more challenging to identify the causal variant in studies of common, polygenic disease. In fact, GWAS often detect non-causal ("tag") variants that are physically close and/or correlated with the true disease variant¹¹. This is a result of linkage disequilibrium (LD), which describes how often variants are acquired together following mating and genetic recombination¹⁶. The lack of causal inference makes biological interpretation, mechanistic understanding, and downstream drug discovery challenging; however, even with the limitations, GWAS results are actively being validated and found to

provide biological and medical insight for a large number of diseases and traits¹¹. Additionally, identifying variants that have a correlated effect on disease risk has the potential to enable the development of strong genetic risk prediction models.

The ability to predict an individual's risk is important, especially for diseases with established screening or prevention strategies; by understanding who is at risk for breast cancer we can better personalize the age at which they receive a mammogram and if we know an individual is at high risk for coronary artery disease, they can be given statins to prevent heart attack¹⁷. Generally, a clinically useful prediction model should be able to stratify individuals by disease risk and accurately predict the probability that a currently asymptomatic individual will eventually develop a disease of interest¹⁸. For example, individuals with a high penetrance risk variant in BRCA1/2 can have up to a 75% lifetime risk of developing breast cancer. As discussed previously, most variants have very small effects on disease risk; therefore, an alternative approach is needed to make accurate risk predictions for common, complex traits. One such approach, shown to have potential clinical utility, is the polygenic risk score (PRS). PRS additively combine many risk variants, with small effects, into a single score with a stronger disease association. For traits such as coronary artery disease, atrial fibrillation, and type 2 diabetes, individuals in the top percentiles of the PRS have been shown to have a greater than 3-fold increased risk of the disease compared to the overall population^{17,19}. Although there are limitations, that will be discussed in detail below (*Section 1.4*), early PRS studies have shown tremendous potential as a clinical instrument for disease detection and prevention especially in combination with known lifestyle and environmental risk factors^{17,19}.

1.3 Cancer genetics and pleiotropy

Cancer is one of the leading causes of death in the world population and nearly 40% of males and females are at risk for developing cancer at some point in their lifetime²⁰; therefore, strategies for early prediction and prevention are necessary to combat the global burden of cancer. GWAS have demonstrated the polygenic architecture of many cancers and identified hundreds of rare and common risk variants that can

be leveraged for risk prediction, as well as for deciphering shared mechanisms of cancer growth²¹. Our prior work in prostate cancer found novel rare variants with evidence of purifying selection at known prostate cancer risk gene, *HOXB13*²². Additionally, by integrating gene expression and protein affinity data, we highlighted the potential of using functional data to link biological mechanisms to noncoding variants associated with prostate cancer risk²². We have also identified pleiotropic variation, or variants that are associated with more than one trait, in cancer. In our study of 18 different cancers we identified 100 pleiotropic associations, providing further evidence for genomic regions that may elucidate key functional pathways activated during cancer development²³. In addition to single-variant pleiotropy, we also illustrated previously unreported pleiotropic patterns using PRS for cross-cancer prediction²⁴. Of the 16 cancer-specific PRSs evaluated, 11 were associated with risk for a separate cancer²⁴. One key component of reducing the global burden of cancer is utilizing pleiotropic variation and risk models to characterize the shared genetic susceptibility and mechanisms of cancer, which will inform screening strategies.

1.4 The lack of diversity in human genetic studies

There have been over 5,700 GWAS conducted covering more than 3,300 traits and identifying greater than 71,000 variant-trait associations¹¹. These studies represent tremendous progress in the field and highlight a clinical potential for utilizing germline variation to inform screening strategies, enable precision medicine, and guide lifestyle factors for disease risk reduction. However, less than 20% of individuals represented in GWAS are of non-European ancestry, which is not reflective of the global representation of ancestries in the world population²⁵. The overrepresentation of Europeans in GWAS is due to a number of factors; however, main contributors are the continued use of existing cohorts, like the UK Biobank which is largely individuals of European ancestry, as well as challenges in recruitment of non-European individuals into genetic studies as a result of, historically supported, mistrust for clinical research²⁶.

The Eurocentric bias in GWAS has severe implications with potential to contribute to health inequity. The majority of current genetic discoveries primarily benefit individuals of European ancestry and do not

replicate in studies of non-European populations^{26,27}. The limited transferability of European findings in non-European populations can be attributed to differences in LD across ancestries between the true causal variant and discovered European tag variant or as a result of population-specific causal disease variants²⁶. Thus, European discovered variation may have less clinical significance in non-European populations. PRS developed in European populations combine many weak trans-ancestry predictors into a single score and the accumulation of these variants leads to an even larger discrepancy in predictive accuracy for diverse, non-European populations. For 17 anthropometric and blood-based traits, there was a 4.5-fold decrease in accuracy when applying European-derived PRSs to individuals of African ancestry compared to Europeans²⁸. The limited utility of European findings in underrepresented populations has been widely demonstrated and although a the majority of studies have been restricted to European populations²⁵⁻²⁹, there is a push to improve diversity and inclusion in genomics research. Many studies are underway to close the gap in genetic discovery for diverse populations, including NIH's All of Us Research Program³⁰, the Human Heredity and Health in Africa (H3Africa)³¹ consortium, and the Population Architecture through Genomics and Environment (PAGE)^{27,32} study, among others.

Expanding genetic studies to non-European populations has immense potential to unlock novel insights into genetic diversity, facilitate fine mapping of causal disease variation, and reduce health disparities. In studies of non-European populations from PAGE, Wojcik *et al.* identified 65 variants that were not previously discovered in studies of European populations that had ~5x the number of samples³². It is also possible to leverage existing large-European studies, while being inclusive and improving power for discovery in underrepresented populations through a trans-ancestry meta-analysis. Trans-ancestry meta-analyses aim to identify potentially causal variants by combining GWAS from multiple ancestries while accounting for potential population differences in linkage disequilibrium and effect estimates³³. It has been shown through recent trans-ancestry studies that the discovery of novel genetic associations is unlikely to result from continuing to increase sample size of European populations, but rather to incorporate studies of diverse populations^{32,34,35}.

In this dissertation, I aim to provide additional evidence supporting the inclusion of diverse populations in genetic studies to enable discovery. Through a simulation study, I will highlight the implications of underrepresentation in GWAS as it relates to the clinical use of PRS and recommend strategies for limiting the disparity in predictive accuracy across populations. I will also provide a real-world example of a trans-ancestry exome-wide association study of multiple primary tumors, a complex cancer phenotype. Overall, my work explores critical areas in genetics research that contribute to the discovery and downstream application of genetic findings in diverse populations.

References

1. International Human Genome Sequencing Consortium *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
2. Venter, J. C. *et al.* The Sequence of the Human Genome. *Science* **291**, 1304–1351 (2001).
3. Weber, J. L. & Myers, E. W. Human Whole-Genome Shotgun Sequencing. *Genome Res.* **7**, 401–409 (1997).
4. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
5. †The International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796 (2003).
6. the Haplotype Reference Consortium. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* **48**, 1279–1283 (2016).
7. NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).
8. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat Rev Genet* **11**, 499–511 (2010).
9. Sudlow, C. *et al.* UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Med* **12**, e1001779 (2015).
10. Hoffmann, T. J. *et al.* Next generation genome-wide association tool: Design and coverage of a high-throughput European-optimized SNP array. *Genomics* **98**, 79–89 (2011).
11. Uffelmann, E. *et al.* Genome-wide association studies. *Nat Rev Methods Primers* **1**, 59 (2021).
12. Badano, J. L. & Katsanis, N. Beyond Mendel: an evolving view of human genetic disease transmission. *Nat Rev Genet* **3**, 779–789 (2002).
13. Rahit, K. M. T. H. & Tarailo-Graovac, M. Genetic Modifiers and Rare Mendelian Disease. *Genes (Basel)* **11**, E239 (2020).
14. Baptista, P. V. Principles in genetic risk assessment. *Ther Clin Risk Manag* **1**, 15–20 (2005).

15. Wray, N. R. *et al.* From Basic Science to Clinical Application of Polygenic Risk Scores: A Primer. *JAMA Psychiatry* **78**, 101 (2021).
16. Slatkin, M. Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. *Nat Rev Genet* **9**, 477–485 (2008).
17. Torkamani, A., Wineinger, N. E. & Topol, E. J. The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics* **19**, 581–590 (2018).
18. Chatterjee, N., Shi, J. & García-Closas, M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nature Reviews Genetics* **17**, 392–406 (2016).
19. Khera, A. V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics* **50**, 1219–1224 (2018).
20. Sasieni, P. D., Shelton, J., Ormiston-Smith, N., Thomson, C. S. & Silcocks, P. B. What is the lifetime risk of developing cancer?: the effect of adjusting for multiple primaries. *Br J Cancer* **105**, 460–465 (2011).
21. Sud, A., Kinnersley, B. & Houlston, R. S. Genome-wide association studies of cancer: current insights and future perspectives. *Nat Rev Cancer* **17**, 692–704 (2017).
22. Emami, N. C. *et al.* A Large-Scale Association Study Detects Novel Rare Variants, Risk Genes, Functional Elements, and Polygenic Architecture of Prostate Cancer Susceptibility. *Cancer Research* **81**, 1695–1703 (2021).
23. Rashkin, S. R. *et al.* Pan-cancer study detects genetic risk variants and shared genetic basis in two large cohorts. *Nat Commun* **11**, 4423 (2020).
24. Graff, R. E. *et al.* Cross-cancer evaluation of polygenic risk scores for 16 cancer types in two large cohorts. *Nat Commun* **12**, 970 (2021).
25. Peterson, R. E. *et al.* Genome-wide Association Studies in Ancestrally Diverse Populations: Opportunities, Methods, Pitfalls, and Recommendations. *Cell* **179**, 589–603 (2019).
26. Sirugo, G., Williams, S. M. & Tishkoff, S. A. The Missing Diversity in Human Genetic Studies. *Cell* **177**, 26–31 (2019).

27. Carlson, C. S. *et al.* Generalization and Dilution of Association Results from European GWAS in Populations of Non-European Ancestry: The PAGE Study. *PLoS Biology* **11**, e1001661 (2013).
28. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature Genetics* **51**, 584 (2019).
29. Bentley, A. R., Callier, S. & Rotimi, C. N. Diversity and inclusion in genomic research: why the uneven progress? *J Community Genet* **8**, 255–266 (2017).
30. Armstrong, K. & Ritchie, C. Research Participation in Marginalized Communities - Overcoming Barriers. *N Engl J Med* **386**, 203–205 (2022).
31. Mulder, N. *et al.* H3Africa: current perspectives. *Pharmgenomics Pers Med* **11**, 59–66 (2018).
32. Wojcik, G. L. *et al.* Genetic analyses of diverse populations improves discovery for complex traits. *Nature* **570**, 514–518 (2019).
33. Li, Y. R. & Keating, B. J. Trans-ethnic genome-wide association studies: advantages and challenges of mapping in diverse populations. *Genome Med* **6**, 91 (2014).
34. Conti, D. V. *et al.* Trans-ancestry genome-wide association meta-analysis of prostate cancer identifies new susceptibility loci and informs genetic risk prediction. *Nat Genet* **53**, 65–75 (2021).
35. Graham, S. E. *et al.* The power of genetic diversity in genome-wide association studies of lipids. *Nature* **600**, 675–679 (2021).

CHAPTER II

Addressing the limitations of polygenic risk scores

2.1 Abstract

The majority of polygenic risk scores (PRS) have been developed and optimized in individuals of European ancestry and may have limited generalizability across other ancestral populations. Understanding aspects of PRS that contribute to this issue and determining solutions is complicated by disease-specific genetic architecture and limited knowledge of sharing of causal variants and effect sizes across populations. Motivated by these challenges, we undertook a simulation study to assess the relationship between ancestry and the potential bias in PRS developed in European ancestry populations. Our simulations show that the magnitude of this bias increases with increasing divergence from European ancestry, and this is attributed to population differences in linkage disequilibrium and allele frequencies of European discovered variants, likely as a result of genetic drift. Importantly, we find that including into the PRS variants discovered in African ancestry individuals has the potential to achieve unbiased estimates of genetic risk across global populations and admixed individuals. We confirm our simulation findings in an analysis of HbA1c, asthma, and prostate cancer in the UK Biobank. Given the demonstrated improvement in PRS prediction accuracy, recruiting larger diverse cohorts will be crucial—and potentially even necessary—for enabling accurate and equitable genetic risk prediction across populations.

2.2 Introduction

Increasing research into polygenic risk scores (PRS) for disease prediction highlights their clinical potential for informing screening, therapeutics, and lifestyle¹. While their use enables risk prediction in individuals of European ancestry, PRS can have widely varying and much lower accuracy when applied to non-European populations²⁻⁴. Although the nature of this bias is not well understood, it can be attributed to the vast overrepresentation of European ancestry individuals in genome-wide association studies (GWAS),

which is 4.5-fold higher than their percentage of the world population; conversely, there is underrepresentation of diverse populations such as individuals of African ancestry in GWAS, which is one fifth their percentage³. Potential explanations for the limited portability of European derived PRS across populations includes differences in population allele frequencies and linkage disequilibrium, the presence of population-specific causal variants or effects, or potential differences in gene-gene or gene-environment interactions⁴. However, in traits such as body mass index and type 2 diabetes, 70 to 80% of European-based PRS accuracy loss in African ancestry has been attributed to differences in allele frequency and linkage disequilibrium; therefore, most causal variants discovered in Europeans are likely to be shared⁵. Recent methods developed to improve PRS accuracy in non-Europeans have prioritized the use of European discovered variants and population specific weighting⁶⁻⁸. However, only small gains in accuracy are possible with limited sample sizes of non-European cohorts⁴.

PRS have been applied and characterized within global populations, but there is limited understanding of PRS accuracy in recently admixed individuals and whether this varies with ancestry. Studies applying PRS in diverse populations^{3-5,9} or exploring potential statistical approaches to improve accuracy in such populations^{6,10} typically present performance metrics averaged across all admixed individuals. Only one study to date has suggested that PRS accuracy may be a function of genetic admixture (i.e., for height in the UK Biobank⁸). However, it is unknown if the relationship between accuracy and ancestry exists when variants are discovered in non-European populations or what the best approach for applying PRS to admixed individuals will be when there are adequately powered GWAS in non-European populations.

To help answer these questions, here we systematically and empirically explore the relationship between PRS performance and ancestry within African, European, and admixed ancestry populations through simulations. We highlight PRS building approaches that will achieve unbiased estimates across global populations and admixed individuals with future recruitment and representation of non-European ancestry individuals in GWAS. We also investigate reasons for loss of PRS accuracy, and attribute this to population

differences in linkage disequilibrium (LD) tagging of causal variants by lead GWAS variants, as well as allele frequency biases potentially due to genetic drift undergone by European ancestry populations. Finally, we confirm our simulation findings by application to data on HbA1c levels, asthma, and prostate cancer in individuals of European and individuals of African ancestry from the UK Biobank.

2.3 Materials and Methods

2.3.1 Simulation of population genotypes

We used the coalescent model (msprime v.7.3¹¹) to simulate European (CEU) and African (YRI) genotypes, based on whole-genome sequencing of HapMap populations, for chromosome 20 as described previously by Martin et al.² Genotypes were modeled after the demographic history of human expansion out of Africa¹², assuming a mutation rate of 2×10^{-8} . We simulated 200,000 Europeans and 200,000 Africans for each simulation trial, for a total of 50 independent simulations (20 million total individuals). We generated founders from an additional 1,000 Europeans and 1,000 Africans (10,000 total across the 50 simulations) to simulate 5,000 admixed individuals (250,000 total across the 50 simulations) with RFMIX v.2¹³ assuming two-way admixture between Europeans and Africans with random mating and 8 generations of admixture.

2.3.2 True and GWAS estimated polygenic risk scores

We generated true genetic liability for all European, African, and admixed individuals within each simulation trial². Briefly, m variants evenly spaced throughout the simulated genotypes were selected to be causal and the effect sizes (β) were drawn from a normal distribution $\beta \sim N\left(0, \frac{h^2}{m}\right)$, where h^2 is the heritability². Constant heritability and complete sharing of effect sizes in African ancestry and European ancestry individuals was assumed. The true genetic liability was computed as the summation of all variant effects multiplied by their genotype for each individual ($X = \sum_{i=1}^m \beta_m g_m$) standardized to ensure total variance of h^2 ($G = \frac{X - \mu_X}{\sigma_X} * \sqrt{h^2}$). Finally, the non-genetic effect ($\varepsilon = N(0, 1 - h^2)$) standardized to

explain the remainder of the phenotypic variation ($E = \frac{\varepsilon - \mu_\varepsilon}{\sigma_\varepsilon} * \sqrt{1 - h^2}$) was added to the genetic risk defining the total trait liability $(G + E)^2$. Cases were selected from the extreme tail of the liability distribution, assuming a 5% disease prevalence. An equal number of controls and 5,000 testing samples were randomly selected from the remainder of the distribution; all 5,000 admixed individuals were also used for testing. Across simulation replicates we varied causal variants ($m = \{200, 500, 1000\}$) and trait heritability ($h^2 = \{0.33, 0.50, 0.67\}$); however, for simplicity main text results assume $m = 1000$ and $h^2 = 0.50$.

The estimated PRS were constructed from GWAS of the simulated genotypes (modeled after chromosome 20) in European and African ancestry populations, each with 10,000 cases and 10,000 controls. Odds ratios (ORs) were estimated for all variants with minor allele frequency (MAF) $> 1\%$ and statistical significance of association was assessed with a chi-squared test. While causal variants may be included in the estimated PRS, they are drawn from the total allele frequency spectrum; thus, they are primarily rare (93% and 87% of causal variants have $MAF < 1\%$ in European and African ancestry populations when $m = 1000$) and restricted from our analysis. For each population, variants were selected for inclusion into the estimated PRS by p-value thresholding ($p = 0.01$ (*Main Text*), 1×10^{-4} , and 1×10^{-6} (*Supplements*)) and clumping ($r^2 < 0.2$) in a 1 Mb window within the GWAS population, where r^2 is the squared Pearson correlation between pairs of variants. A fixed-effects meta-analysis was also performed to calculate the inverse-variance weighted average of the ORs in African and European ancestry populations, and LD r^2 values for clumping used both datasets as the reference.

For each individual, an estimated PRS was calculated as the sum of the $\log(\text{OR})$ (i.e., the PRS ‘weights’) multiplied by their genotype for all independent and significant variants at a given threshold. The PRS were constructed for testing samples with variants and weights each selected from European or African ancestry GWAS, or a fixed-effects meta of both combined. Additional multi-ancestry PRS approaches^{7,10} were also

explored for admixed individuals. Accuracy was measured by Pearson’s correlation (r) between the true genetic liability and estimated PRS within each population. Across simulation trials, averages and ninety-five percent confidence intervals for r were calculated following a Fisher z-transformation for approximate normality¹⁴. The statistical significance of differences in accuracy between PRS approaches was assessed within ancestry groups, defined by global genome-wide European ancestry proportions, with a z-test (also following Fisher transformation). Specifically, within each simulation trial the z-statistic, measuring the difference between two PRS approaches, was computed and a two-sided p-value was obtained; results were summarized across trials by taking the median p-value. While using r as a measure of accuracy has the added benefit of being independent from heritability, in admixed individuals we also calculate the proportion of variance (R^2) for total trait liability (genetic and environmental component) explained by the estimated PRS.

2.3.3 *Multi-ancestry polygenic risk scores*

Local Ancestry Weighting PRS

In addition to genotypes of simulated admixed individuals, RFMIX¹³ also outputs the local ancestry at each locus for every individual. Thus, we used this information to create a local ancestry weighted PRS that is not impacted by ancestry inference errors. For admixed African and European ancestry individuals an ancestry-specific PRS was constructed for each population (k) by limiting each PRS to variants found in that ancestry-specific subset of the genome ($i \in k$), as defined by local ancestry, and weighting using variant effects discovered in that population⁷. Each ancestry-specific PRS was then combined, weighted by the genome-wide global ancestry proportion (ρ_k) for that individual as follows⁷:

$$PRS = \rho_{EUR} \sum_{i \in EUR} \beta_{i,EUR} G_i + (1 - \rho_{EUR}) \sum_{i \in AFR} \beta_{i,AFR} G_i$$

In this way each individual has a PRS constructed from the same independent variants with personalized weights that are unique to the individual’s local ancestry.

Linear Mixture of Multiple Ancestry-Specific PRS

Using a linear mixture approach developed by Márquez-Luna et al.¹⁰ we combined two PRS estimated in each of our global training populations

$$PRS = \alpha_1 PRS_{EUR} + \alpha_2 PRS_{AFR}$$

where individual PRS were constructed using significant and independent variants ($p < 0.01$ and $r^2 < 0.2$ in a 1Mb window) and effect sizes from a GWAS in that training population. For simulations, mixing weights (α_1 and α_2) were estimated in an independent African ancestry testing population and as validation, accuracy was assessed in our simulated admixed ancestry individuals.

2.3.4 Application to real data

We obtained genome-wide summary statistics for HbA1c¹⁵, asthma^{16,17}, and prostate cancer^{18,19} calculated in European and African ancestry individuals (Table S1). Summary statistic variants that were not present in both the UK Biobank European and African ancestry testing populations were removed. PRS for each phenotype were constructed from associated and independent GWAS variants within each training population by p-value thresholding ($p = \{5 \times 10^{-8}, 1 \times 10^{-7}, 5 \times 10^{-7}, 1 \times 10^{-6}, 5 \times 10^{-6}, 1 \times 10^{-5}, 5 \times 10^{-5}, 1 \times 10^{-4}, 5 \times 10^{-4}, 1 \times 10^{-3}, 5 \times 10^{-3}, 0.01, 0.05, 0.1, 0.5, 1\}$) and clumping ($LD\ r^2 < 0.2$) of variants within 1Mb with PLINK²⁰. Additionally a fixed-effects meta-analysis of the two populations was performed using METASOFT v2.0.1²¹. The selected PRS variants exhibited limited heterogeneity between the European and African ancestry training set summary statistics. In particular, of all possible European (African) ancestry selected PRS variants, only 5.4% (9.4%), 6.9% (5.7%), and 7.0% (4.8%) were heterogeneous between the two groups for HbA1c, asthma, and prostate cancer, respectively (i.e., $I^2 > 25\%$ and $Q\ p\text{-value} < 0.05$).

PRS performance was evaluated in an independent cohort using genotype and phenotype data for individuals of European ancestry and individuals of African ancestry (Table S1) from the UK Biobank, imputation and quality control previously described²². We undertook extensive post-imputation quality control of the UK Biobank, including the exclusion of relatives and ancestral outliers from within each

group. Specifically, analyses were limited to self-reported European and African ancestry individuals, with additional samples excluded if genetic ancestry PCs did not fall within five standard deviations of the self-reported population mean. For each individual, their PRS was computed as the weighted sum of the genotype estimates of effect on each phenotype from the discovery studies (Table S1), multiplied by the genotype at each variant. For each population-specific variant set, weights from either the European or African summary statistics or the fixed-effects meta-analysis were used. A total of 96 polygenic risk scores were evaluated in each phenotype exploring the impact of ancestral population (two scenarios), p-value threshold (16 scenarios), and variant weighting (three scenarios). The proportion of variation explained by each PRS (partial- R^2) approach was assessed for UKB European-ancestry and African-ancestry individuals separately. The partial- R^2 was calculated from the difference in R^2 values following linear regression of HbA1c levels on age, sex, BMI, and PCs (1-10) with and without the PRS also included. Similarly, for asthma and prostate cancer, we determined the Nagelkerke's pseudo partial- R^2 following logistic regression of case status on age, sex (asthma only), BMI (prostate cancer only), and PCs (1-10) with and without the PRS. Additionally, in African ancestry individuals we created a combined PRS ($\alpha_1 PRS_{EUR} + \alpha_2 PRS_{AFR}$) where PRS_{EUR} and PRS_{AFR} was the most optimal PRS using variants from the designated population and the weight and p-value that resulted in the highest accuracy; albeit in sample, optimization was done within a single PRS to ensure limited overfitting of the combined model¹⁰. We used 5-fold cross validation to assess model performance in which 80% of the cohort was used to estimate the mixing coefficients (α_1 and α_2) and the combined PRS partial- R^2 was calculated in the remaining 20% of the data. Reported partial- R^2 was averaged across folds¹⁰. For our binary phenotypes with unbalanced cases and controls we used stratified 5-fold cross validation.

2.4 Results

2.4.1 Generalizability of European-derived risk scores to an admixed population

We constructed PRS from our simulated European datasets and applied them to independent simulated European, African, and admixed testing populations, assuming 1000 true causal variants (m) and trait

heritability (h^2) of 0.5. On average, 1552 (range = [1134-1920]) variants were selected for inclusion into the PRS at p -value < 0.01 and LD $r^2 < 0.2$ (Table 1). The average accuracy across replicates (50 simulations), measured by the correlation (r) between the true and inferred genetic risk, was much higher when applying the PRS to Europeans ($r = 0.77$; 95% CI = [0.76, 0.77]) than to Africans ($r = 0.45$; 95% CI = [0.44, 0.47]; Figure 1). This is in agreement with decreased performance seen in real data when applying a European derived genetic risk score to an African population²⁻⁵.

To understand the relationship between ancestry and PRS accuracy, admixed individuals were stratified by their proportion of genome-wide European (CEU) ancestry: high (100%>CEU>80%), intermediate (80%>CEU>20%), and low (20%>CEU>0%). PRS performance decreased with decreasing European ancestry (Figure 1). Average accuracy (Pearson's correlation) for the high, intermediate, and low European ancestry groups was 0.73 (95% CI = [0.72, 0.74]), 0.61 (95% CI = [0.60, 0.62]), and 0.53 (95% CI = [0.51, 0.54]), respectively (Figure 1). In comparison to Europeans, the performance of the European derived PRS was significantly lower in individuals with intermediate (20% decrease, $p = 1.27 \times 10^{-47}$), and low (32% decrease, $p = 6.48 \times 10^{-16}$) European ancestry, and with African-only ancestry (41% decrease, $p = 8.00 \times 10^{-155}$). There was no significant difference for individuals with high (5.3% decrease, $p = 0.09$) European ancestry. These trends remained consistent when varying the genetic architecture (Figure S1), specifically decreasing the number of causal variants (m) and varying the trait heritability (h^2). Additionally, the relationship between ancestry and accuracy persisted with the inclusion of variants at lower p -value thresholds (Figure S2).

By further binning admixed individuals into deciles of global European ancestry and determining the variance explained of the total liability (genetics and environment) by the PRS, we estimated a 1.34% increase in accuracy for each 10% increase in global European ancestry, replicating a previous analysis of height in the UK Biobank⁸. The slope of this linear relationship increased with increasing heritability but was not found to vary with the number of true causal variants (Figure S3).

2.4.2 Population-specific weighting of European selected variants

Using a well-powered GWAS from our simulated African cohort (10,000 cases and 10,000 controls), we aimed to explore the potential accuracy gains achieved from a PRS with European selected variants, but with population specific weighting of these variants. We applied three different weighting approaches to incorporate non-European effect sizes: (1) effect sizes from an African ancestry GWAS for all variants; (2) effect sizes from a fixed-effects meta-analysis of European and African ancestry GWAS for all variants, both having 10,000 cases and 10,000 controls; and (3) population specific weights based on the local ancestry for an individual at each variant in the PRS (Figure 2).

The most accurate PRS approach varied by the proportion of European ancestry. Populations with greater than 20% African ancestry benefited significantly from the inclusion of population specific weights (Figure 2). Intermediate European ancestry benefitted most from using fixed-effects meta-analysis weighting instead of European weights ($r = 0.64$ vs. 0.61 , $p = 0.02$). In contrast, variant weighting from an African ancestry GWAS instead of from European had higher accuracy in low European ancestry ($r = 0.65$ vs. 0.53 , $p = 0.009$) and African-only ($r = 0.64$ vs. 0.45 , $p = 2.02 \times 10^{-44}$) populations. Individuals with high European ancestry had similar accuracy with weights from a fixed-effects meta-analysis as from European ($r = 0.73$ in both, $p = 0.79$), but decreased performance with the inclusion of weights from an African ancestry GWAS ($r = 0.62$ vs. 0.73 , $p = 0.01$).

No clear benefits, and in some cases significant decreases, were observed for local ancestry informed weights compared to weights from a European or African ancestry GWAS or fixed-effects meta-analysis. Individuals with high, intermediate, and low European ancestry had lower accuracy using local ancestry informed weights compared to the best weighting in each ancestry group: $r = 0.71$ vs. 0.73 (from fixed-effect or European weights; $p = 0.58$); $r = 0.61$ vs. 0.64 (from fixed-effect weights; $p = 0.004$); and $r = 0.63$ vs. 0.65 (from African weights; $p = 0.60$), respectively (Figure 2).

2.4.3 Performance of non-European PRS variant selection and weighting approaches

In our simulations, population specific weighting of PRS SNPs discovered in European ancestry populations improved PRS accuracy; however, the disparity between performance in European ancestry individuals versus African and admixed ancestry individuals remained large. We aimed to explore the potential improvements in PRS that could be gained by including variants discovered in large, adequately powered African ancestry cohorts. Following clumping and thresholding of significant variants using LD and summary statistics from the simulated African populations, an average of 5269 (range = [4462-6071]) variants were included in the PRS (Table 1) reflective of the greater genetic diversity and decreased LD compared to Europeans²³. In contrast, when constructing a PRS using the same LD and p-value criteria applied to a fixed-effects meta-analysis of European and African ancestry, an average of only 92 (range = [38-197]) variants were included in the PRS. This substantially smaller number was a result of few variants being statistically significant in both populations. Of the total number of variants included from the European GWAS, African ancestry GWAS, and fixed-effects meta, only 1.15%, 0.54%, and 15.0% on average were the exact causal variant from the simulation; an additional 3.72%, 5.34%, and 33.3% tagged at least one causal variant with $r^2 > 0.2$ (and were within ± 1000 kb of that causal variant) in European ancestry populations and 3.45%, 2.42%, and 28.1% in African ancestry populations (Table 1). The limited overlap between true causal and GWAS selected variants is a result of causal variants in our simulation arising from the total spectrum of allele frequencies, and therefore more likely to be rare, while GWAS is better powered to detect common variants in the study population from which they were identified². These common variants may not adequately tag rare variants due to low correlation²⁴.

Overall, we constructed twelve PRS with variants selected from GWAS in European or African ancestry populations or a fixed-effects meta of both (three scenarios) and weights from the same approaches plus an additional local ancestry specific weighting method (four scenarios) (Figure 2). For Europeans, the highest PRS accuracy was achieved with European selected variants and weights ($r = 0.77$; 95% CI = [0.76, 0.77]);

however, a similar accuracy was observed for weights from a fixed-effects meta ($r = 0.76$; $p = 0.53$). For Africans, the highest PRS accuracy was with African selected variants and weights from a fixed-effects meta ($r = 0.75$; 95% CI = [0.74, 0.75]), similar performance was observed with African variants and weights ($r = 0.74$, $p = 0.28$). For admixed individuals, the highest performing PRS depended on the population ancestry percentage. In individuals with high European ancestry (>80%), the best PRS was with European selected variants and fixed-effects meta or European weights ($r = 0.73$; 95% CI = [0.72, 0.74]). For individuals with intermediate (20%-80%) or low (<20%) European ancestry, the most accurate PRS was from using African selected variants and weights from a fixed-effects meta-analysis ($r = 0.68$; 95% CI = [0.67, 0.68] and 0.71 ; 95% CI = [0.70, 0.72], respectively). Again, no benefit was observed with the inclusion of local ancestry specific weights for any set of PRS variants. Using a more stringent p-value threshold and including fewer variants into the PRS did not result in a change of the best PRS variants and weights (Figure S2).

2.4.4 Inclusion of variants from diverse populations

We found that including in the PRS variants discovered in African ancestry GWAS with population specific weights results in less disparity in PRS accuracy across ancestries compared to European selected variants, confirming that GWAS in non-bottlenecked populations may yield a more unbiased set of disease variants²⁵. For example, applying to individuals of African ancestry a PRS derived from GWAS variants and weights discovered in training data from the target population results in a 15.7% higher accuracy compared to using a PRS comprised of variants discovered in a European GWAS (also with African weights). In contrast, the gains in accuracy achieved by sourcing variants from ancestry-matched studies were much lower in European ancestry individuals. Compared to a PRS with variants from an African ancestry GWAS (with European weights), a PRS derived from a European GWAS (also with European weights) only gave a 3.9% higher accuracy. We also observed better generalization of PRS based on African selected variants across all admixed groups (Figure 2).

Unlike in Europeans, a PRS with variants discovered in African ancestry populations generalized across ancestral groups with population-specific weighting. However, similar to the European PRS, the African ancestry derived PRS (with African variants and weights) was estimated to have a 1.62% increase in the variance explained of the total trait liability by the PRS for each 10% increase in African ancestry (Figure S4). Through a linear combination of the European and African ancestry derived PRS (Methods)¹⁰, the relationship between ancestry and accuracy diminished to less than a 0.4% increase per 10% increase of African ancestry (Figure S4).

While the best single PRS for admixed individuals with at least 20% African ancestry selected variants based on a GWAS in an African ancestry population with weights from a fixed-effects meta-analysis, a linear combination of the European and African ancestry derived PRS had higher accuracy; this was particularly true at decreased African ancestry cohort sizes. We saw considerable improvements with the combined PRS over using a European derived (European selected variants and weights) PRS, especially for low European ancestry (CEU < 20%) where even with 10-fold fewer African samples there was a 27.4% increase in PRS accuracy compared to the European derived risk score and a 12.3% increase compared to a PRS with African ancestry selected variants and weights from a fixed-effects meta (Figure 3).

2.4.5 Allele frequency and LD of GWAS variants

We sought to understand what factors impacted PRS generalizability across the different variant selection approaches. GWAS performed in European and African ancestry populations (for SNPs with MAF \geq 0.01) were both more likely to identify significant variants that were more common in their own population than in the other. Approximately 60% of variants identified in European ancestry populations had minor allele frequencies less than 1% in African ancestry populations and vice-versa; however, given the underlying assumption of homogeneity, the smaller number of variants selected by a meta-analysis of the two populations tended to have more similar minor allele frequencies (Figure 4a). Although European and African ancestry GWAS were both better powered to detect variants at intermediate frequencies within the

same study population, GWAS in European ancestry populations may be unable to capture derived risk variants that have remained in Africa, which could be the result of genetic drift, whereas GWAS in African ancestry populations are not subject to this bias²⁵.

We also examined LD tagging of causal variants by GWAS selected variants within our simulated European and African populations. Each causal variant's LD score was calculated by summing up the LD r^2 between that causal variant and every GWAS tag variant within ± 1000 kb. The LD scores calculated in European and African ancestry populations were highly correlated (Pearson's $r > 0.7$) for the GWAS and fixed-effects meta selected variants. Variants selected from a fixed-effects meta had the highest LD score correlation between populations, as expected given that the variants reached significance in both populations and therefore were more common with similar LD patterns (Figure 4b). Since LD score correlation did not vary largely between simulations, we examined the raw LD scores for a single simulation in order to illustrate differences in LD score magnitude not captured by the Pearson's correlation.

We found that European selected variants had higher LD scores in European compared to in African ancestry populations; however, variants selected from an African ancestry GWAS tagged causal variants in both populations more strongly (Figure 4c). This is unlikely to be due to the larger number of African selected variants, as the results were unchanged following normalization of LD scores by dividing the total LD score for each causal variant by PRS size (Figure S5). Fixed-effects meta-analysis variants had similar LD score magnitudes. However, while a greater proportion of the fixed-effects meta selected variants were causal, fewer were strong tags for causal variants not directly identified, highlighting the potential need for a model that does not assume homogeneity of effects for tag variants²⁶. Additionally, causal variants with identical effect sizes may have differing allele frequencies across populations which would result in heterogeneous allele substitution effects; this helps indicate why a fixed-effects meta-analysis may not be the optimal approach.

2.4.6 Application to real data

To corroborate our simulation findings, we undertook an analysis of 96 PRS developed for the prediction of multiple complex traits in European and African ancestry individuals from the UK Biobank, including HbA1c levels, asthma status, and prostate cancer (Table S1). We tested variant selection strategies based on p-value thresholding and LD clumping of genome-wide summary statistics¹⁵ computed in independent European or African ancestry cohorts and variant weights from the same approaches with an additional weighting from a fixed-effects meta across populations. Multiple p-value thresholds and weighting strategies were tested to assess the PRS accuracy in African ancestry individuals relative to European ancestry individuals across parameters.

In UK Biobank Europeans, a GWAS significant European-derived PRS (with European variants and weights) had a partial- R^2 of 1.6%, 1.2%, and 1.5% respectively for HbA1c levels, asthma, and prostate cancer; the same PRS applied to African ancestry individuals, with approximately 13.1% European ancestry⁸, only explained 0.07%, 0.38%, and 0.19% (Figure S6). Although the proportion of variation explained by the PRS (partial- R^2) was consistently lower in UK Biobank African ancestry individuals compared to Europeans, prediction was improved through the inclusion of variants or weights discovered in an independent African ancestry cohort across all traits (Figure S6). Interestingly, we found that a linear combination of the best performing PRS with European discovered variants and African ancestry discovered variants improved accuracy substantially (Table S2), supporting our simulation finding that a combined PRS which includes variants from the target population, even at smaller sample sizes, is the optimal approach for constructing PRS in admixed and non-European individuals.

2.5 Discussion

Our work shows that incorporating variants selected from European GWAS into a PRS can result in less accurate prediction in non-European and admixed populations in comparison to variants selected from a well-powered African ancestry GWAS. Through simulations and application to real data analysis of

multiple complex traits, we provide empirical evidence that supports the use of a linear mixture of multiple population derived PRS to remove bias with ancestry and achieve higher accuracy in admixed individuals with currently available non-European sample sizes. We also demonstrate the anticipated improvements in PRS prediction accuracy that may be achieved with the inclusion of diverse individuals in GWAS, highlighting the need to actively recruit non-European populations.

Our simulation finding that prediction accuracy of a European derived PRS linearly decreases with increasing proportion of African ancestry in admixed African and European populations is consistent with a recent study of height where there was a 1.3% decrease for each 10% increase in African ancestry⁸. This decrease in prediction accuracy has been attributed to linkage disequilibrium and allele frequency differences, as well as differences in effect sizes across populations contributing to height⁸. Our work adds further insights into this reduction in PRS accuracy, showing that (1) it exists in the absence of trans-ancestry effect size differences consistent with previous theoretical models that did look at admixture^{2,5}, and (2) variants selected from an African population may not have these same biases. Recent work found that known GWAS loci discovered in Europeans have allele frequencies that are upwardly biased by 1.15% in African ancestry populations which results in a misestimated PRS; a phenomenon that likely arises due to population bottlenecks and ascertainment bias from GWAS arrays²⁵. In our simulation study, which was not impacted by ascertainment bias, we show that GWAS in African ancestry populations also identify variants with population allele frequency differences; however, these variants have more consistent LD tagging of causal variants across populations. Our observations support the hypothesis that well-powered African ancestry GWAS will be able to more accurately capture disease associated loci that are more broadly applicable across populations, due to having undergone less genetic drift²⁵.

A major advantage of our study is having large simulated European and African ancestry cohorts to provide guidelines for developing the best possible PRS in admixed individuals with current and future GWAS. Through our exploration of 12 PRS, with various variant selection and weighting approaches, we re-

capitulate recent results applying similar PRS strategies to an admixed Hispanic/Latino population⁹. For individuals with intermediate proportions of European ancestry (20-80%), we also see improvements using European selected variants and population-specific or fixed-effects meta weights; however, as non-European cohorts get increasingly large it will be imperative to perform variant discovery in these populations as gains in accuracy with weight adjustment of European selected variants will be limited especially in individuals with higher proportions of non-European ancestry.

Current methods for improving PRS accuracy in diverse populations have prioritized the inclusion of variants from European GWAS, as these have higher statistical power to identify trait associated loci. For example, one such approach uses a two-component linear mixed model to allow for the incorporation of ethnic-specific weights⁶. Another method creates ancestry-specific partial PRS for each individual based on the local ancestry of variants selected from a European GWAS⁷. This approach was found to improve trait predictability, compared to a traditional PRS with population specific or European weights, in East Asians for BMI but not height⁷. In contrast, implementing this local-ancestry method⁷ in our simulation, we found that PRS accuracy was higher with African or fixed-effects meta weighting than with local ancestry in admixed African ancestry populations. Our results suggest that true equality in performance will be difficult to obtain solely based on European-identified variants even with local ancestry-adjusted weights. Although local ancestry weighting may have greater benefits when complete sharing across populations is not assumed, we show that in the absence of population-specific factors, the optimal PRS approach involves using variants identified in a large African population and population-specific weighting.

To create a multi-ancestry PRS without incorporating local ancestry, *Márquez-Luna et al. (2017)* uses a mixture of PRS taking advantage of existing well-powered GWAS studies and supplementing with additional information that can be gained from a smaller study in the population of interest¹⁰. While this approach may offer relative improvement in PRS accuracy for non-Europeans compared to a European-derived PRS, our simulation suggests that the inclusion of significant tag variants discovered in Europeans

may unnecessarily hinder predictive performance in non-Europeans. We investigate this approach in the context of varying admixture proportions and find that it achieved high accuracy across all admixed individuals, was not biased by ancestry, and significantly improved performance over a European-only PRS with 10-fold fewer African ancestry cases. Thus, a combination of multiple single population PRS may be the best currently available approach for admixed individuals, and this approach will likely continue to improve as the individual PRS are further developed.

An important novel finding of our work that the inclusion of variants from an African-ancestry population results in less disparity in PRS accuracy across other populations, illustrates the need to recruit diverse populations in GWAS and make these data readily available. Large consortia such as H3Africa, PAGE, the Million Veterans Program, and All of Us are undertaking efforts to support this initiative. Based on our analysis of HbA1c, asthma, and prostate cancer in the UK Biobank, we find that improvement in PRS prediction accuracy is currently possible by incorporating findings from GWAS in African ancestry populations, albeit with lower power. In addition to smaller sample sizes, this potential improvement may be limited by ascertainment bias in what SNPs are included on genotyping arrays and poorer imputation in non-Europeans. GWAS arrays and their imputation have substantially higher coverage among Europeans, and this may result in decreased PRS portability of findings across traits; in such situations, whole genome sequencing in diverse populations may be needed in order to improve accuracy^{27,28}. Our study and others support the immense genetic diversity that can be unlocked by studying underrepresented populations to both eliminate the disparity in genetics for prediction medicine and provide novel insights into disease biology for all populations^{25,27,29}.

Although our simulation study provides important insight into the future of PRS use, it has important limitations. First, while coalescent simulations allow for decreased computational burden, model assumptions may result in inaccurate long-range linkage disequilibrium especially for whole genome simulations³⁰. However, given we only simulated chromosome 20, biases are expected to be modest³⁰. We

also use a case-control framework for our simulation; therefore, power and potential differences in population PRS accuracy may be even higher if a quantitative trait was used. In addition, our simulations assume random mating among admixed individuals and therefore do not reflect the more complex assortative mating that may be observed, which may impact the distribution of local ancestry tract lengths in our simulation and therefore hinder the improvement of PRS accuracy by local ancestry weighting³¹. Also, although we provide evidence to suggest the contribution of population differences in allele frequency and LD tagging of causal variants to loss of PRS accuracy with varying ancestry, we do not delineate how each of these factors decrease accuracy independently; this is a direction for future work. Finally, we have only simulated individuals from Yoruba, a West African population, which is not representative of the greater diversity in Sub Saharan Africa³². Future studies must be done to ensure our findings can be extended to individuals from other regions of Africa.

Overall, our findings support the concern that while studies have demonstrated the potential clinical utility of PRS, adopting the current versions of these scores could contribute to inequality in healthcare⁴. Going forward, future studies should prioritize the inclusion of diverse participants and care must be taken with the interpretation of currently available risk scores. While statistical approaches may offer improvements in accuracy compared to current European-derived risk scores, in order to truly diminish the disparity and achieve PRS accuracies similar to in European ancestry populations we must actively recruit and study diverse populations.

2.6 Acknowledgements

This material is based on work supported by the National Science Foundation Graduate Research Fellowship Program under grant No. 1650113 and NIH grant No. CA201358. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. This research has been conducted using the UK

Biobank Resource under application number 14015. Furthermore, the authors thank Linda Kachuri for providing helpful feedback and discussion.

2.7 Tables

Table 2.1 Summary of PRS variants and causal tagging across simulations

Table 2.1 Legend: The set of PRS variants from each GWAS and the fixed-effects meta-analysis were selected by p-value thresholding ($p < 0.01$) and clumping ($r^2 < 0.2$) across the 50 simulations. Each PRS variant was compared to the causal set of variants ($m = 1000$) within each simulation to determine the direct overlap between the two sets and the LD r^2 between the PRS variant and every causal variant within a 1000 kb window. The total number of selected PRS variants that tag at least one causal variant at r^2 greater than 0.8, 0.6, 0.4, or 0.2 is listed in the table.

GWAS Population	Total # PRS Variants ($p < 0.01$)	# Causal	# in LD with a Causal Variant			
			$r^2 > 0.8$	$r^2 > 0.6$	$r^2 > 0.4$	$r^2 > 0.2$
European	1552 [1134-1920]	18 [10-26]				
LD in Europeans			27 [16-40]	32 [22-44]	39 [25-55]	58 [38-80]
LD in Africans			20 [9-36]	25 [16-42]	34 [24-54]	53 [35-70]
African	5269 [4462-6071]	28 [18-40]	-	-	-	-
LD in Europeans			94 [67-122]	132 [95-171]	183 [123-238]	280 [202-364]
LD in Africans			37 [26-48]	48 [34-61]	67 [50-89]	127 [81-170]
Fixed-Effects Meta	92 [38-197]	12 [5-22]	-	-	-	-
LD in Europeans			15 [6-26]	17 [6-28]	21 [9-39]	29 [16-47]
LD in Africans			13 [6-21]	14 [6-25]	17 [9-29]	24 [10-43]

* The number of variants is reported as the average and range [low-high] across the 50 simulations.

2.8 Figures

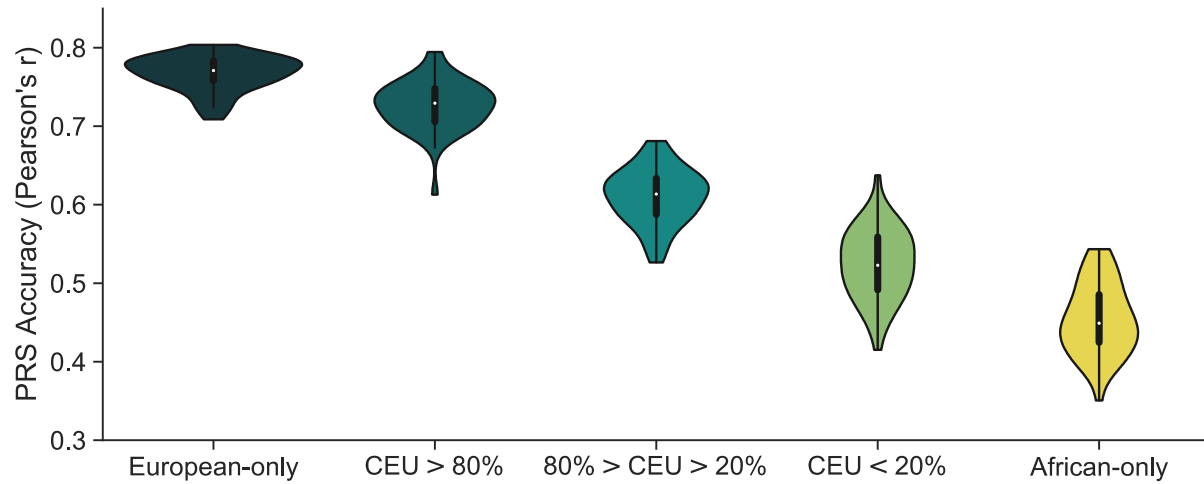


Figure 2.1 Accuracy of European derived PRSs by proportion of total ancestry

Figure 2.1 Legend: Variants and weights were extracted from a GWAS of 10000 European cases and 10000 European controls. PRS accuracy was computed as the Pearson's correlation between the true genetic risk and GWAS estimated risk score across 50 simulations in independent test populations of 5000 Europeans, 5000 Africans, and 5000 admixed individuals. Admixed individuals were grouped based on their proportion of genome-wide European ancestry. Simulations assume 1000 causal variants and a heritability of 0.5 to compute the true genetic risk. A p-value of 0.01 and LD r^2 cutoff of 0.2 was used to select variants for the estimated risk score.

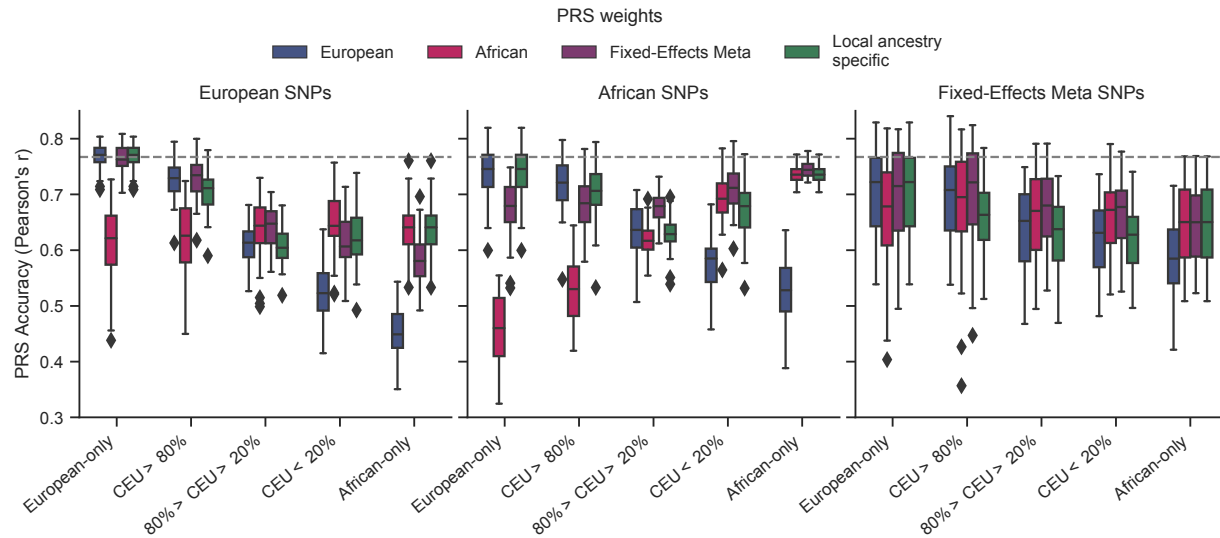


Figure 2.2 PRS construction approaches and performance in admixed individuals

Figure 2.2 Legend: PRS were constructed using variants and weights selected from either a European or African population (10000 cases, 10000 controls each) or a fixed-effects meta-analysis of both. An additional local ancestry specific method was used for PRS weighting. Performance, measured as the Pearson's correlation between the true and GWAS estimated risk score, is shown across 50 simulations. Simulations assume 1000 causal variants and a heritability of 0.5 to compute the true genetic risk. A p-value of 0.01 and LD r^2 cutoff of 0.2 was used to select variants for the estimated risk scores.

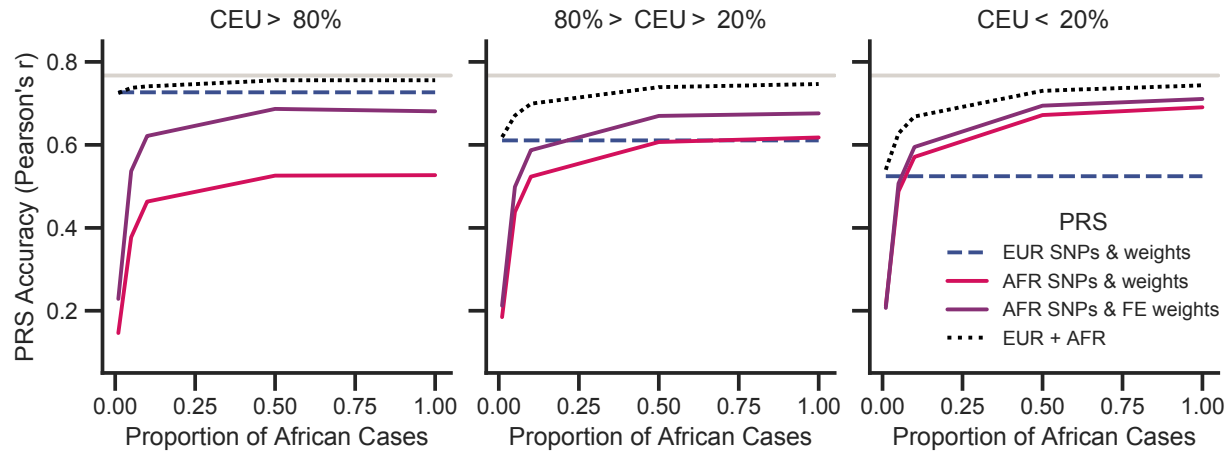


Figure 2.3 Impact of African ancestry sample size on PRS accuracy and generalization

Figure 2.3 Legend: The number of African samples used in the GWAS and subsequent PRS construction was decreased to reflect availability of diverse samples in real data. Analysis was conducted assuming 1%, 5%, 10%, 50%, and 100% (matched size of European dataset) of the total African ancestry cases. Average accuracy and the 95% confidence interval were reported across the 50 simulations for different variant selection and weighting approaches. Simulations assume 1000 causal variants and a heritability of 0.5 to compute the true genetic risk. A p-value of 0.01 and LD r^2 cutoff of 0.2 was used to select variants for the estimated risk score. A linear mixture of single population PRS ($\alpha_1 EUR + \alpha_2 AFR$), with variants and weights selected from that population, was also tested in the admixed population. The mixture coefficients (α_1 and α_2) were estimated in an independent African ancestry testing population.

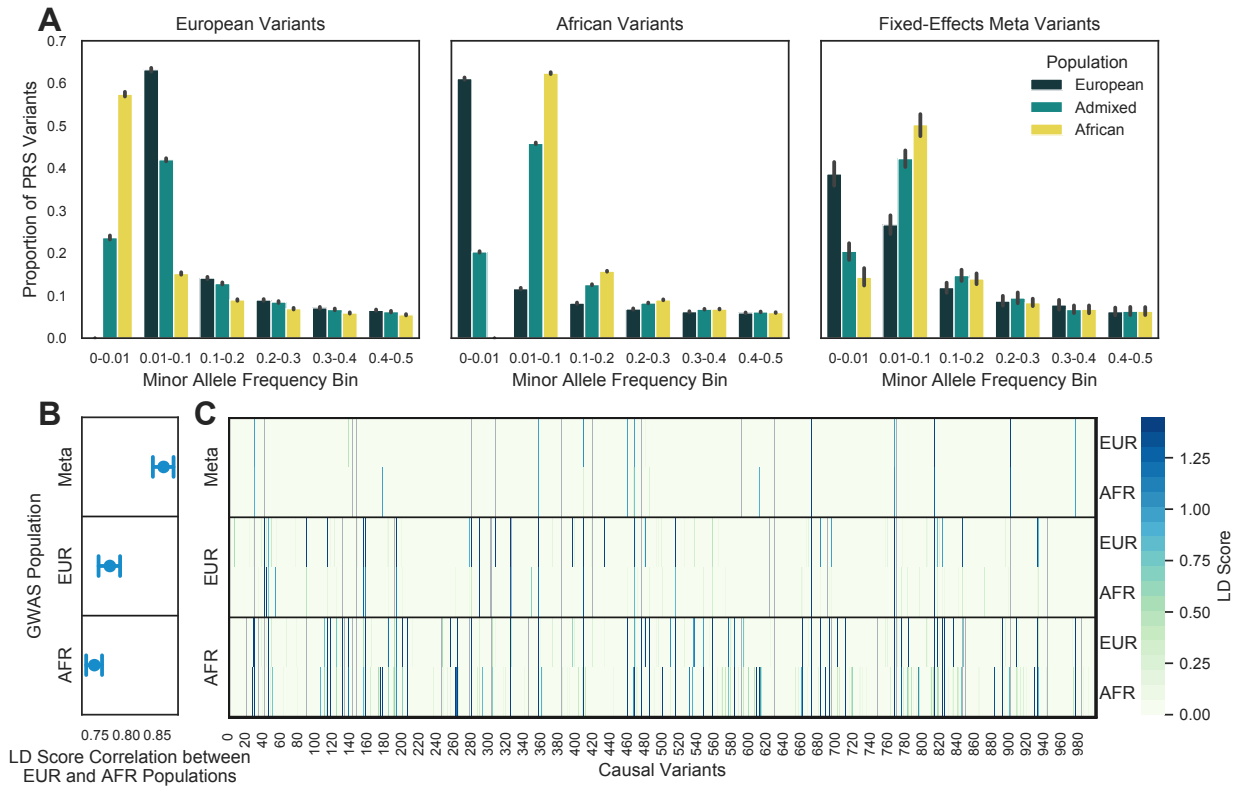


Figure 2.4 Allele frequency distribution of GWAS selected variants and LD tagging of causal variants

Figure 2.4 Legend: Variants were selected from a European or African ancestry GWAS or a fixed-effects meta of both populations. 4a. GWAS variants were binned by their minor allele frequency estimated from the European, African, and admixed populations. The error bar represents the 95% CI across simulations. 4b. LD scores were calculated for every causal variant by adding up the LD r^2 for each GWAS tag variant within ± 1000 kb of the causal variant. LD scores calculated in a Europeans and Africans were compared by Pearson's correlation. The results were summarized across simulations as the average and 95% CI. 4c. Raw LD scores for each causal variant ($m = 1000$) calculated in a European or African population for one simulation. Each panel shows the approach used for variant selection. Causal variants directly discovered through the GWAS are colored in grey.

2.9 Supplementary Materials

Table S2.1 Summary of HbA1c variants included for each PRS

Table S2.1 Legend: Using summary statistics for HbA1c from Wheeler et. al. 2017¹⁵, variants were selected for inclusion into the PRS following p-value thresholding and LD clumping ($r^2 < 0.2$) within 250kb windows. The intersection between PRS variants selected from either a European or African GWAS and available sites in UKB Europeans and Africans is reported.

PRS Training Dataset ^a	P-Value	N Variants Available Testing / N Variants in Score (%)	
		UKB Europeans	UKB Africans
European	p < 5e-8	83 / 84 (98.81%)	64 / 84 (76.19%)
	p < 1e-7	87 / 88 (98.86%)	66 / 88 (75.0%)
	p < 5e-7	107 / 108 (99.07%)	82 / 108 (75.93%)
	p < 1e-6	118 / 119 (99.16%)	91 / 119 (76.47%)
	p < 5e-6	158 / 159 (99.37%)	121 / 159 (76.1%)
	p < 1e-5	187 / 188 (99.47%)	141 / 188 (75.0%)
	p < 5e-5	274 / 275 (99.64%)	206 / 275 (74.91%)
	p < 1e-4	354 / 355 (99.72%)	257 / 355 (72.39%)
	p < 5e-4	724 / 726 (99.72%)	490 / 726 (67.49%)
	p < 1e-3	1,113 / 1,117 (99.64%)	726 / 1,117 (65.0%)
	p < 5e-3	3,498 / 3,503 (99.86%)	2,169 / 3,503 (61.92%)
	p < 0.01	6,006 / 6,016 (99.83%)	3,705 / 6,016 (61.59%)
	p < 0.05	22,917 / 22,943 (99.89%)	13,772 / 22,943 (60.03%)
	p < 0.1	40,980 / 41,040 (99.85%)	24,821 / 41,040 (60.48%)
	p < 0.5	137,803 / 137,990 (99.86%)	85,660 / 137,990 (62.08%)
p < 1	195,825 / 196,161 (99.83%)	125,585 / 196,161 (64.02%)	
African	p < 5e-8 ^b	1 / 1 (100.0%)	1 / 1 (100.0%)
	p < 1e-7 ^b	1 / 1 (100.0%)	1 / 1 (100.0%)
	p < 5e-7	2 / 2 (100.0%)	2 / 2 (100.0%)
	p < 1e-6	2 / 2 (100.0%)	2 / 2 (100.0%)
	p < 5e-6	10 / 10 (100.0%)	10 / 10 (100.0%)
	p < 1e-5	16 / 16 (100.0%)	16 / 16 (100.0%)
	p < 5e-5	60 / 61 (98.36%)	61 / 61 (100.0%)
	p < 1e-4	109 / 111 (98.2%)	111 / 111 (100.0%)
	p < 5e-4	464 / 472 (98.31%)	471 / 472 (99.79%)
	p < 1e-3	874 / 891 (98.09%)	887 / 891 (99.55%)
	p < 5e-3	3,906 / 3,988 (97.94%)	3,975 / 3,988 (99.67%)
	p < 0.01	7,294 / 7,429 (98.18%)	7,406 / 7,429 (99.69%)
	p < 0.05	30,575 / 31,129 (98.22%)	31,025 / 31,129 (99.67%)
	p < 0.1	55,859 / 56,940 (98.1%)	56,687 / 56,940 (99.56%)
	p < 0.5	210,251 / 214,648 (97.95%)	213,662 / 214,648 (99.54%)
p < 1	318,848 / 326,687 (97.6%)	324,706 / 326,687 (99.39%)	

^a GWAS summary statistics extracted from Wheeler et. al. 2017.

^b PRS is not computed when only one variant is available.

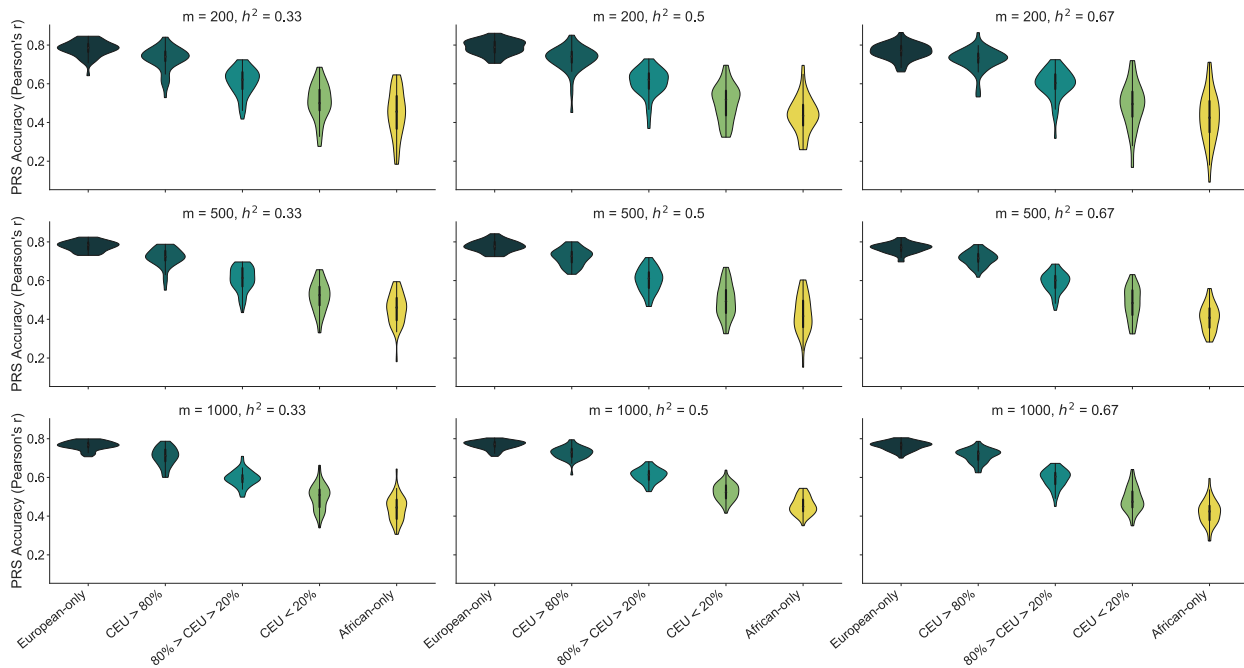


Figure S2.1 European derived risk score accuracy with varying simulation causal variants and assumed heritability

Figure S2.1 Legend: Our simulation assumes a set number of causal variants (m) and trait heritability (h^2) when generating the true genetic risk score. We varied these parameters and tested all combinations assuming $m = \{200, 500, 1000\}$ and $h^2 = \{0.33, 0.5, 0.67\}$. The accuracy was measured by Pearson's correlation between the true and GWAS estimated risk score for each of the 50 simulations. We used a European GWAS to select independent variants and effect sizes for the PRS ($p < 0.01$ and LD $r^2 = 0.2$). This risk score was applied to Europeans, Africans, and admixed populations.

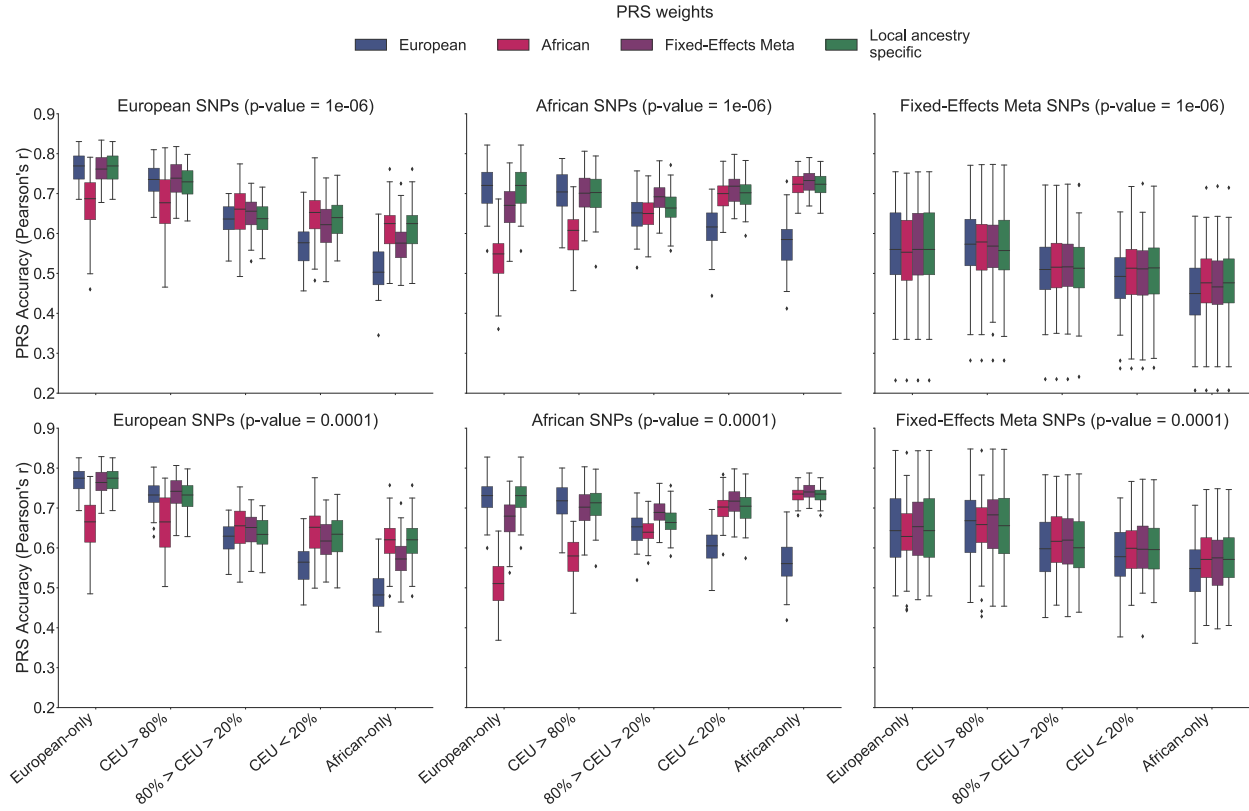


Figure S2.2 PRS accuracy with varying p-value thresholds

Figure S2.2 Legend: The p-value used for variant selection approach was decreased to allow fewer variants into the PRS. In addition to $p < 0.01$, shown in the main text results, we also tested $p < 1 \times 10^{-4}$ and 1×10^{-6} . Simulations assume 1000 causal variants and a heritability of 0.5 to compute the true genetic risk. A p-value of 0.01 and LD r^2 cutoff of 0.2 was used to select variants for the estimated risk scores.

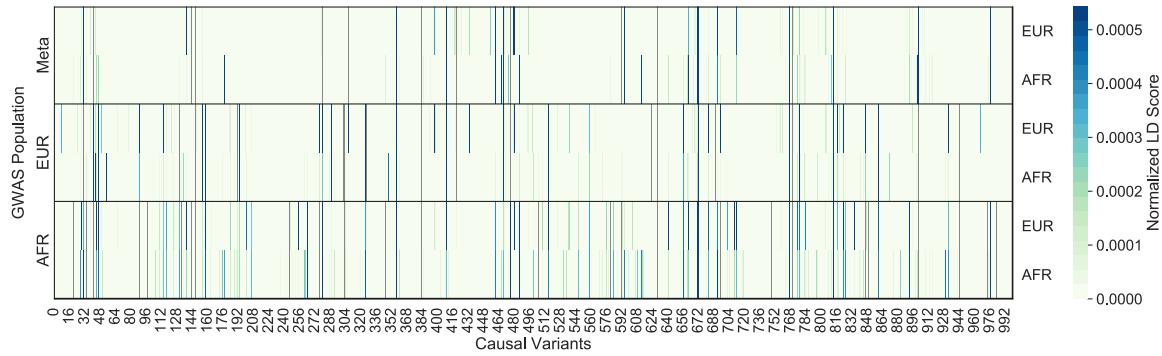


Figure S2.3 Normalized LD tagging of causal variants by GWAS selected variants

Figure S2.3 Legend: LD scores were calculated for every causal variant by adding up the LD r^2 for each GWAS tag variant within ± 1000 kb of the causal variant. LD scores were then normalized by the total number of GWAS selected variants from each population reflected by the 3 panels. The normalized LD scores for each causal variant ($m = 1000$) calculated in a European or African population is shown for one simulation. Causal variants directly discovered through the GWAS are colored in grey.

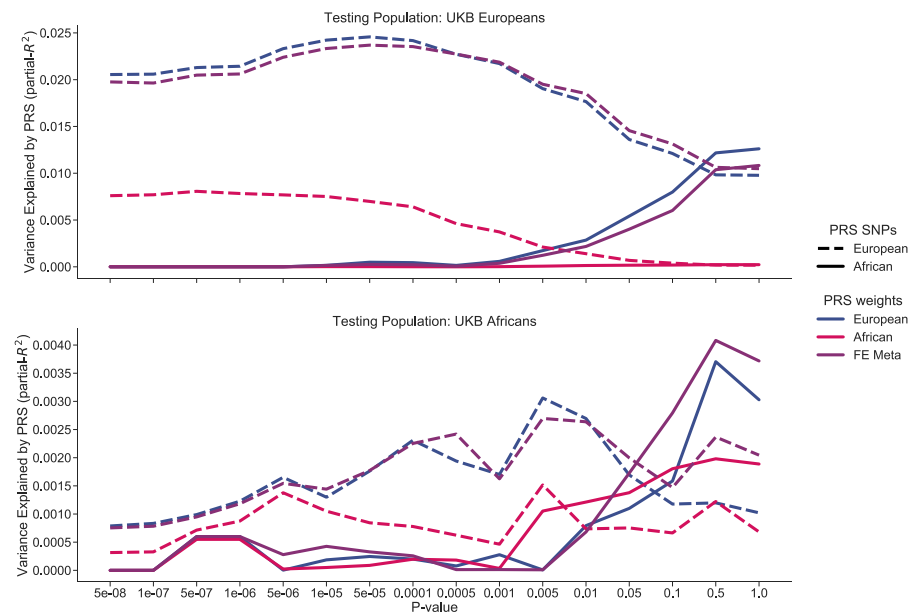


Figure S2.4 HbA1c PRS performance in the UK Biobank

Figure S2.4 Legend: Partial- R^2 was calculated for each PRS, within Europeans and Africans from the UK Biobank, by fitting a linear model of HbA1c with age and sex and subtracting the null R^2 from a model including the PRS in addition to age and sex.

References

1. Torkamani, A., Wineinger, N. E. & Topol, E. J. The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics* **19**, 581–590 (2018).
2. Martin, A. R. *et al.* Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *Am. J. Hum. Genet.* **100**, 635–649 (2017).
3. Duncan, L. *et al.* Analysis of polygenic risk score usage and performance in diverse human populations. *Nat Commun* **10**, 3328 (2019).
4. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature Genetics* **51**, 584 (2019).
5. Wang, Y. *et al.* Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations. *Nat Commun* **11**, 3865 (2020).
6. Coram, M. A., Fang, H., Candille, S. I., Assimes, T. L. & Tang, H. Leveraging Multi-ethnic Evidence for Risk Assessment of Quantitative Traits in Minority Populations. *The American Journal of Human Genetics* **101**, 218–226 (2017).
7. Marnetto, D. *et al.* Ancestry deconvolution and partial polygenic score can improve susceptibility predictions in recently admixed individuals. *Nat Commun* **11**, 1628 (2020).
8. Bitarello, B. D. & Mathieson, I. Polygenic Scores for Height in Admixed Populations. *G3* g3.401658.2020 (2020) doi:10.1534/g3.120.401658.
9. Grinde, K. E. *et al.* Generalizing polygenic risk scores from Europeans to Hispanics/Latinos. *Genetic Epidemiology* **0**,
10. Márquez-Luna, C., Loh, P.-R. & Price, A. L. Multi-ethnic polygenic risk scores improve risk prediction in diverse populations. *Genet Epidemiol* **41**, 811–823 (2017).
11. Kelleher, J., Etheridge, A. M. & McVean, G. Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLoS Comput Biol* **12**, e1004842 (2016).

12. Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H. & Bustamante, C. D. Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLoS Genet* **5**, e1000695 (2009).
13. Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference. *The American Journal of Human Genetics* **93**, 278–288 (2013).
14. Silver, N. C. & Dunlap, W. P. Averaging correlation coefficients: Should Fisher's z transformation be used? *Journal of Applied Psychology* **72**, 146–148 (1987).
15. Wheeler, E. *et al.* Impact of common genetic determinants of Hemoglobin A1c on type 2 diabetes risk and diagnosis in ancestrally diverse populations: A transethnic genome-wide meta-analysis. *PLoS Med* **14**, e1002383 (2017).
16. CAAPA *et al.* Association study in African-admixed populations across the Americas recapitulates asthma risk loci in non-African populations. *Nat Commun* **10**, 880 (2019).
17. Australian Asthma Genetics Consortium (AAGC) collaborators *et al.* Multiancestry association study identifies new asthma risk loci that colocalize with immune-cell enhancer marks. *Nat Genet* **50**, 42–53 (2018).
18. Emami, N. C. *et al.* Association Study of Over 200,000 Subjects Detects Novel Rare Variants, Functional Elements, and Polygenic Architecture of Prostate Cancer Susceptibility. <http://biorxiv.org/lookup/doi/10.1101/2020.02.12.929463> (2020) doi:10.1101/2020.02.12.929463.
19. Conti, D. V. & *et al.* Multiethnic GWAS meta-analysis identifies novel variants and informs genetic risk prediction for prostate cancer across populations. *Nature Genetics* (2020).
20. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics* **81**, 559–575 (2007).
21. Han, B. & Eskin, E. Random-Effects Model Aimed at Discovering Associations in Meta-Analysis of Genome-wide Association Studies. *The American Journal of Human Genetics* **88**, 586–598 (2011).

22. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
23. Campbell, M. C. & Tishkoff, S. A. African Genetic Diversity: Implications for Human Demographic History, Modern Human Origins, and Complex Disease Mapping. *Annu. Rev. Genom. Hum. Genet.* **9**, 403–433 (2008).
24. Li, B. & Leal, S. M. Methods for Detecting Associations with Rare Variants for Common Diseases: Application to Analysis of Sequence Data. *The American Journal of Human Genetics* **83**, 311–321 (2008).
25. Kim, M. S., Patel, K. P., Teng, A. K., Berens, A. J. & Lachance, J. Genetic disease risks can be misestimated across global populations. *Genome Biol* **19**, (2018).
26. Morris, A. P. Transethnic meta-analysis of genomewide association studies: Transethnic Meta-Analysis of GWAS. *Genet. Epidemiol.* **35**, 809–822 (2011).
27. Martin, A. R. *et al.* Low-coverage sequencing cost-effectively detects known and novel variation in underrepresented populations. <http://biorxiv.org/lookup/doi/10.1101/2020.04.27.064832> (2020) doi:10.1101/2020.04.27.064832.
28. Li, J. H., Mazur, C. A., Berisa, T. & Pickrell, J. K. Low-pass sequencing increases the power of GWAS and decreases measurement error of polygenic risk scores compared to genotyping arrays. <http://biorxiv.org/lookup/doi/10.1101/2020.04.29.068452> (2020) doi:10.1101/2020.04.29.068452.
29. Bentley, A. R., Callier, S. L. & Rotimi, C. N. Evaluating the promise of inclusion of African ancestry populations in genomics. *npj Genom. Med.* **5**, 5 (2020).
30. Nelson, D. *et al.* Accounting for long-range correlations in genome-wide simulations of large cohorts. *PLoS Genet* **16**, e1008619 (2020).
31. Zaitlen, N. *et al.* The Effects of Migration and Assortative Mating on Admixture Linkage Disequilibrium. *Genetics* **205**, 375–383 (2017).
32. Gurdasani, D. *et al.* The African Genome Variation Project shapes medical genetics in Africa. *Nature* **517**, 327–332 (2015).

CHAPTER III

Genetic susceptibility to multiple primary tumors

3.1 Abstract

Background: Up to one of every six individuals diagnosed with one cancer will be diagnosed with a second primary cancer in their lifetime. Genetic factors contributing to the development of multiple primary cancers, beyond known cancer syndromes, have been underexplored. **Methods:** To characterize genetic susceptibility to multiple cancers, we conducted a pan-cancer, whole-exome sequencing study of individuals drawn from two large multi-ancestry populations (6,429 cases, 165,853 controls). We created two groupings of individuals diagnosed with multiple primary cancers: 1) an overall combined set with at least two cancers across any of 36 organ sites; and 2) cancer-specific sets defined by an index cancer at one of 16 organ sites with at least 50 cases from each study population. We then investigated whether variants identified from exome sequencing were associated with these sets of multiple cancer cases in comparison to individuals with one and, separately, no cancers. **Results:** We identified 22 variant-phenotype associations, 10 of which have not been previously discovered and were significantly overrepresented among individuals with multiple cancers, compared to those with a single cancer. **Conclusions:** Overall, we describe variants and genes that may play a fundamental role in the development of multiple primary cancers and improve our understanding of shared mechanisms underlying carcinogenesis.

3.2 Introduction

The substantial global burden of cancer coupled with increasing survival due to improved screening, surveillance, and treatments has yielded a growing number of cancer survivors who are at risk of developing a second primary cancer in their lifetime^{1,2}. The prevalence of multiple primary cancers globally is estimated to be between 2 and 17%, with the wide range likely due to differences in cancer registration practices, case definitions, population characteristics, and follow-up times^{1,2}. Cancer predisposition

syndromes, such as Li-Fraumeni, Lynch, and hereditary breast and ovarian cancer, are known to increase the risk of multiple primary cancers; however, less than 2% of all cancers are attributed to hereditary cancer syndromes¹. Genetic risk factors for multiple primary cancers beyond known syndromes are not well understood.

Genome-wide association studies (GWAS) have implicated many common, low penetrance variants in 5p15 (*TERT-CLPTMIL*)³, 6p21 (*HLA*)^{4,5}, 8q24⁶, and other loci in the risk of several cancer types. Additional studies have investigated pleiotropy in these regions or characterized cross-cancer susceptibility variants^{7,8}. A pleiotropic locus has the potential to not only affect risk of many different cancer types, but also increase the likelihood that a single individual develops multiple primary cancers. In our prior work, we discovered that the rare pleiotropic variant *HOXB13* G84E had a stronger association with the risk of developing multiple primary cancers than of a single cancer⁹. This suggests that there may be increased power to detect pleiotropic variation in individuals with multiple primary cancers relative to those with only a single cancer. Identifying widespread pleiotropic signals is informative for understanding shared genetic mechanisms of carcinogenesis, toward the identification of informative markers for cancer prevention and precision medicine.

In this study, we survey the landscape of rare and common variation in individuals with multiple primary cancers, single cancers, and cancer-free controls through whole-exome sequencing (WES) in two large, multi-ancestry studies. We evaluate associations previously discovered in studies of individuals with a single cancer and find novel pleiotropic variation in individuals with multiple primaries.

3.3 Materials and Methods

3.3.1 Study populations and phenotyping

Our study included ancestrally diverse individuals with multiple primary cancers or no cancer from two large study populations: the Kaiser Permanente Research Bank (KPRB) and the UK Biobank (UKB). From

the KPRB, we included individuals who were previously genotyped through the Research Program on Genes, Environment and Health (RPGEH) and the ProHealth Study. For the UKB, we specifically studied participants from the 200K release of WES data, which also included individuals diagnosed with a single cancer¹⁰.

For both study populations, ascertainment of cancer diagnoses has been previously described^{7,11}. Both studies included prevalent and/or incident diagnoses of malignant, borderline, and in situ primary tumors¹¹. ICD codes indicating non-melanoma skin cancer or metastatic cancer were not considered primary tumors. Cancers were primarily defined according to the SEER site recode paradigm¹². However, for hematologic cancers, we incorporated morphology following WHO classifications¹³, placing cancers into three major subtypes: lymphoid neoplasms, myeloid neoplasms, and NK- and T-cell neoplasms (Table S1). Cases were individuals with ICD-9 or ICD-10 codes for primary tumors at two or more distinct organ sites. In the KPRB, controls without a cancer diagnosis were matched 1:1 to cases on age at specimen collection, sex, genotyping array, and reagent kit. In the UKB, controls included all individuals without a cancer diagnosis.

In both study populations, we excluded duplicates/twins and first-degree relatives, retaining the individual from each related pair who had higher coverage at targeted sites. Following quality control (QC) of WES data (described below), the KPRB and UKB study populations used in this project included 3,111 and 3,318 cases with multiple primary cancers and 3,136 and 162,717 cancer-free controls, respectively. The UKB also contributed 29,091 individuals with a single cancer diagnosis. While our study was primarily unselected for cancer type, prostate cancer cases were oversampled in the KPRB due to inclusion of individuals from the ProHealth Study.

3.3.2 Genetic ancestry and principal components analysis

Genetic ancestry was defined using genome-wide, imputed array data that underwent extensive QC, as previously described¹¹. Ancestry principal components (PCs) were computed using flashPCA2¹⁴ by

projecting our study samples onto PCs defined by 1000G phase 3 reference populations¹⁵. Individuals were assigned to the closest reference population using distance from the top 10 PCs. Individuals with ancestral PCs greater than five standard deviations from the reference population mean were excluded. The final analytic dataset included individuals of European, African, East Asian, South Asian, and Hispanic/Latino ancestry; however, the analysis was largely biased towards individuals of European ancestry as they were overrepresented (Figure S1). A total of $N = 646$ (10.2%) and $N = 8,739$ (5.26%) individuals were of non-European ancestry in the KPRB and UKB, respectively (Table 1).

3.3.3 Whole-exome sequencing and quality control

The Regeneron Genetics Center used the Illumina NovaSeq 6000 platform to perform WES for both study populations where the source of DNA was saliva for the KPRB and blood for the UKB. Sample preparation and QC were performed using a high-throughput, fully-automated process that has been previously described in detail¹⁶. Briefly, following sequencing, reads were aligned to the GRCh38 reference genome and variants were called with WeCall¹⁶ for the KPRB and DeepVariant¹⁷ for the UKB. Samples with gender discordance, 20x coverage at less than 80% of targeted sites, and/or contamination greater than 5% were excluded.

Additional QC was applied to filter low quality variants and related individuals. First, genotype calls with low depth of coverage (DP) were updated to missing ($DP < 7$ for SNPs and $DP < 10$ for indels). Then, sites with low allele balance (AB) were removed. Specifically, variants without at least one sample having $AB \geq 15\%$ for SNPs or $AB \geq 20\%$ for indels were excluded. Following previous studies¹⁶, we excluded variants with missingness $> 10\%$ and HWE p-value $< 10^{-15}$, computed across all individuals in each study population. After these steps, a total of ~ 3.51 M high-quality sites were retained for the KPRB and ~ 15.92 M were retained for the UKB; excluding singletons, there were ~ 1.36 M and ~ 8.22 M variants, respectively. In the UKB, the larger number of variants observed was due to rare variation present in the larger sample size;

when restricting to common variants (MAF > 1%), there were ~186K and ~137K variants, respectively for the KPRB and UKB.

3.3.4 Association analysis in individuals with multiple cancers versus cancer-free controls

Genetic association analyses of single variants and genes investigated the following cancer phenotypes: (1) diagnosis with at least two primary cancers across any of the 36 organ sites ("any 2+ primary cancers") and (2) groupings of individuals defined by a shared index cancer at one of 16 organ sites with at least 50 cases from each study population ("cancer-specific analyses"). Primary analyses compared multiple cancer cases to cancer-free controls. Within our cancer-specific analyses of 16 organ sites, there were cases shared across our index cancer groupings. For example, the set of individuals with at least one diagnosis of breast cancer overlaps with those having at least one ovarian cancer diagnosis.

Single-variant and gene-based association analyses were performed using REGENIE v2.2.4, a machine-learning approach for performing whole-genome regression to correct for cryptic population structure, as well as, adjust for case-control imbalance by applying saddlepoint approximation when the standard case-control p-value is less than 0.05¹⁸. We assessed single-variant associations for high-quality variants shared across both populations with minor allele count (MAC) > 2 across cancer phenotype cases and controls within each study. The number of variants tested in our single-variant analyses varied by cancer phenotype (~337K [other female genital cancer-specific analysis] to ~722K [any 2+ primary cancers]). WES variants were functionally annotated using SnpEff v5.0¹⁹ and dbNFSP v3.5²⁰ accessed through ANNOVAR²¹. Missense variants were classified using five algorithms: (1) SIFT ("D"); (2) HDIV from Polyphen2; (3) HVAR from Polyphen2; (4) LRT ("D"); and (5) MutationTaster ("A" or "D"). For our gene-based burden analyses, we restricted to rare variants with a MAF < 0.5%, including singletons, computed across all individuals within each study population. Following previous work, three gene-based models were evaluated and the model with the lowest p-value was selected²²: (1) all rare variants with predicted loss-of-function (pLOF) by SnpEff, (2) pLOF and missense rare variants predicted to be deleterious by the above

five classification algorithms, and (3) pLOF and missense rare variants predicted to be deleterious by at least one algorithm. In our gene-based and single-variant analyses, we adjusted for covariates including age, top 10 PCs, and sex (except for sex-specific index cancers of the breast, cervix, ovary, uterus, other female genital organ, and prostate). In the KPRB population, we additionally adjusted for genotyping array and reagent kit, as they were used to perform case-control matching. In the UKB, we adjusted for flow cell (S2 vs S4), which differed for the initial 50K and subsequent 150K release of WES samples.

Single-variant and gene-based burden analyses for each phenotype were combined across study populations in a fixed-effects meta-analysis using METASOFT²³ and metafor v3.0.2²⁴, respectively. For our single-variant analyses, we report all suggestive, independent [linkage disequilibrium (LD) $r^2 < 0.2$] associations with $p < 5 \times 10^{-6}$. For our gene-based analyses, we report all associations adjusted for the number of genes tested ($p < 2.65 \times 10^{-6} = 0.05 / 18,842$). In both analyses, we report meta-analysis p-values.

3.3.5 Distinguishing susceptibility signals for multiple cancers versus single cancers

We also evaluated whether the variants and genes associated with the diagnosis of multiple primary cancers (versus non-cancer controls) remained associated when comparing individuals with multiple cancers to those diagnosed with a single cancer. These analyses assessed whether the variants or genes were pleiotropic for developing multiple cancers or general markers of susceptibility to a specific cancer. We undertook these analyses in the UKB sample only, since individuals diagnosed with a single primary cancer were not sequenced in the KPRB. Single-variant and gene-level analyses were implemented as described above. For each variant or gene of interest identified in our case-control analyses, we performed a case-case analysis comparing individuals diagnosed with multiple cancers to those diagnosed with a single cancer. For our cancer-specific analyses, we compared individuals diagnosed with the index cancer plus any other cancer to those diagnosed with the index cancer only. For example, for a finding discovered in our cancer-specific analysis of prostate cancer, we performed a case-case analysis comparing individuals diagnosed with prostate cancer plus any other cancer to individuals with only a prostate cancer diagnosis.

3.4 Results

3.4.1. *Characterization of multiple primary cancer diagnoses in two large study populations*

Our meta-analyses included 6,429 cases with multiple primary cancers and 165,853 cancer-free controls (Table 1). All cases had at least two independent primary cancer diagnoses, and 656 cases had more than two diagnoses (Figure S2). In the KPRB, the maximum number of cancer diagnoses for an individual was 6 ($n = 1$) and in the UKB, the maximum number was 5 ($n = 2$). Overall, 36 unique cancer sites were represented across multiple cancer cases in the two study populations, with 180 unique pairs of sites (e.g., breast and melanoma) and 298 unique ordered pairs of sites by diagnostic sequence (e.g., breast followed by melanoma) (Table S2). Only 51 of the 298 ordered pairs had at least 25 cancer cases when grouping individuals by first and second cancer diagnosis (i.e., ignoring any subsequent cancer diagnoses; Table S2, Figure 1). The top ordered pairs represented in the combined study populations were prostate then melanoma ($N = 221$), cervix then breast ($N = 202$), melanoma then prostate ($N = 180$), breast then melanoma ($N = 174$), and prostate then colorectal ($N = 170$). Prostate, breast, melanoma, colorectal, and cervix were the most common sites of first cancer diagnoses (Figure 1). The prevalence of each cancer pair was similar in the KPRB and UKB (Figure S3). As most individual cancer pairs were underpowered for downstream analysis, we considered all multi-cancer cases combined, as well as groupings of individuals with a shared index cancer (16 cancers) (Figure S4, Table S3). Among those with multiple cancers, the cancers with the largest number of cases were prostate ($N = 1,977$; oversampled in KPRB), breast ($N = 1,874$), melanoma ($N = 1,443$), colorectal ($N = 1,324$), and urinary bladder ($N = 829$).

3.4.2. *Exome-wide single variant association analyses*

We found two independent, genome-wide significant associations ($p < 5 \times 10^{-8}$) and 20 suggestive associations ($p < 5 \times 10^{-6}$) between individual variants and the multiple cancer phenotypes (i.e., either any 2+ primary cancers or cancer-specific analyses) (Figure 2, Table S4). We found an additional two significant and two suggestive associations (Figure S5) in our cancer-specific analyses of lymphoid and

myeloid neoplasms; however, we assumed them to represent somatic alterations in the blood as they had low allele balance across our heterogeneous samples (Figure S6) and occur in genes known to be impacted by clonal hematopoiesis of indeterminate potential (CHIP)²⁵. Results were relatively homogeneous across the KPRB and UKB study populations (Table S4).

Of our 22 findings, two variants were suggestively associated with any 2+ primary cancers, rs555607708 (OR [95% CI] = 2.72 [1.79, 4.15], $p = 3.10 \times 10^{-6}$), a frameshift variant in *CHEK2* known to be associated with risk at many cancer sites²⁶, and rs146381257 (OR [95% CI] = 7.82 [3.28, 18.62], $p = 3.45 \times 10^{-6}$), a 5'upstream variant in *ZNF106*. The risk-increasing allele for rs555607708 (*CHEK2*) was most commonly found among individuals with at least one breast cancer (41.9%), prostate cancer (30.6%), melanoma (22.6%), or cervical cancer (16.1%) (Figure 2). For rs146381257 (*ZNF106*), frequencies were increased in prostate cancer (33.3%), lung cancer (28.6%), breast cancer (28.6%), lymphoid neoplasms (23.8%), urinary bladder cancer (19.0%), pancreatic cancer (14.3%), and kidney cancer (14.3%).

An additional 10 of our findings were previously reported risk variants for a single cancer (Figure 2). Notably, we detected an association with the *MC1R* variant rs1805008 for melanoma²⁷ (OR [95% CI] = 1.56 [1.35, 1.81], $p = 2.73 \times 10^{-9}$), when comparing all individuals with at least one melanoma diagnosis plus any other cancer diagnosis to cancer-free controls. We also replicated the previously associated prostate-specific antigen (PSA) variant, rs17632542²⁸ (*KLK3*, OR [95% CI] = 1.49 [1.28, 1.73], $p = 3.87 \times 10^{-7}$) in individuals with at least one prostate cancer diagnosis. In addition, we replicated associations between missense risk variant rs6998061 (8q24 locus, *POU5F1B*) and multiple tumor types in both our prostate cancer-specific analysis²⁹ (OR [95% CI] = 1.23 [1.13, 1.33], $p = 4.39 \times 10^{-7}$) and our colorectal cancer-specific analysis³⁰ (OR [95% CI] = 1.25 [1.15, 1.37], $p = 1.06 \times 10^{-7}$).

The remaining variants demonstrating associations with multiple cancer phenotypes were not previously associated with any single cancer (Figure 2). They included a variant discovered in our breast cancer-

specific analysis, rs143745791 (*NCBPI*, OR [95% CI] = 5.95 [2.79, 12.67], $p = 3.76 \times 10^{-6}$), for which 16.2% of carriers, restricted to cases, had a breast and cervical cancer diagnosis, and a variant discovered in our urinary bladder cancer-specific analysis, rs141647689 (*SDKI*, OR [95% CI] = 9.29 [3.63, 23.80], $p = 3.45 \times 10^{-6}$), for which 14.3% of carriers also had prostate cancer (Figure 2). Three variants found in our lymphoid neoplasm-specific analysis had increased frequencies in cases who also had a diagnosis of prostate cancer: rs535484207 (*RANBP2*, OR [95% CI] = 256.01 [26.82, 2,442.95], $p = 1.46 \times 10^{-6}$), rs139586367 (*UFLI*, OR [95% CI] = 284.06 [27.95, 2,886.15], $p = 1.79 \times 10^{-6}$), and rs191064896 (*ADGRB1*, OR [95% CI] = 108.36 [15.02, 781.08], $p = 3.32 \times 10^{-6}$), where 21.4%, 40.0%, and 25.0% of carriers for the risk-increasing allele, for each respective variant, had both cancers. The *ADGRB1* variant was also present at increased frequencies among individuals with a lymphoid neoplasm and breast cancer diagnosis (25.0%, Figure 2).

3.4.3. Gene-based analyses of multiple cancers

Out of 18,842 genes tested, we found 10 significant associations ($p < 2.65 \times 10^{-6}$) across our analyses of any 2+ primary cancers and our cancer-specific analyses (Figure 3, Table S5). An additional four CHIP genes (*ASXL1*, *TET2*, *JAK2*, and *DDX41*) were significantly associated with myeloid neoplasms and are likely driven by somatic alterations (Figure S7).

In our analyses of any 2+ primary cancers and our breast cancer-specific analysis, we replicated associations for known pleiotropic genes, *BRCA2* (pLOF, $p = 3.76 \times 10^{-11}$ and 1.91×10^{-9}) and *CHEK2* (pLOF + missense, $p = 2.95 \times 10^{-11}$ and 1.67×10^{-8}) (Figure 3). *BRCA2* also emerged in our ovarian cancer-specific analysis (pLOF, $p = 1.91 \times 10^{-9}$). We found associations between the known prostate cancer gene *ATM* and any 2+ primary cancers and in our prostate cancer-specific analysis (pLOF + missense, $p = 9.84 \times 10^{-7}$ and 2.56×10^{-6}). Additional associations were observed between *SAMHD1* and *SLC642* and any 2+ primary cancers (pLOF + missense, $p = 2.40 \times 10^{-7}$ and $p = 5.44 \times 10^{-7}$, respectively). *BRCA1* also surfaced in the breast cancer-specific analysis (pLOF, $p = 6.68 \times 10^{-8}$).

Predicted loss of function variants in *BRCA1* and *BRCA2* were present at increased frequencies in individuals with a breast cancer diagnosis and ovary as an additional cancer site (Figure 3), such that 28.6% and 13.6% of individuals, respectively, were a carrier for at least one variant in the burden set. For *BRCA1*, there was also an increase of carriers with an additional melanoma (9.52%) or lung cancer (9.52%) diagnosis. For *BRCA2*, there was an increase of carriers with an additional uterine (8.47%), lung (6.78%), or colorectal cancer (6.78%).

3.4.4. Comparison of mutation burden in individuals with multiple versus single cancers

Out of the 22 associated variants (Figure 2), 10 remained associated when comparing individuals with multiple cancers to those with single cancers (Table S6; $p < 0.05$). Two of these variants were positively associated in our analysis of any 2+ primary cancers: rs555607708 (*CHEK2*; OR [95% CI] = 1.57 [1.09, 2.25], $p = 0.015$) and rs146381257 (*ZNF106*; OR [95% CI] = 5.38 [1.07, 27.18], $p = 0.042$). The other eight variants were positively associated with diagnosis of a specific index cancer plus any other cancer versus the specific cancer alone (Table S6). Two of these eight variants were associated in our breast cancer-specific case-case analysis: rs7872034, a missense variant in *SMC2* (OR [95% CI] = 1.16 [1.05, 1.27], $p = 0.0025$) and rs143745791, a missense variant in *NCBPI* (OR [95% CI] = 3.71 [2.08, 6.61], $p = 8.37 \times 10^{-6}$).

Of the 10 findings from the gene-level burden analyses (Figure 3), eight remained positively associated with multiple cancers in comparison with single cancers ($p < 0.05$; Table S7). Five of these genes were discovered in our case-case analysis of any 2+ primary cancers: *SLC6A2* (OR [95% CI] = 1.86 [1.42, 2.41], $p = 3.90 \times 10^{-6}$), *ATM* (OR [95% CI] = 1.42 [1.15, 1.77], $p = 1.10 \times 10^{-3}$), *CHEK2* (OR [95% CI] = 1.56 [1.23, 1.98], $p = 2.31 \times 10^{-4}$), *SAMHD1* (OR [95% CI] = 1.56 [1.14, 2.13], $p = 5.34 \times 10^{-3}$), and *BRCA2* (OR [95% CI] = 1.86 [1.31, 2.65], $p = 5.42 \times 10^{-4}$). *ATM* (OR [95% CI] = 1.82 [1.20, 2.75], $p = 4.64 \times 10^{-3}$) was positively associated in our prostate cancer-specific case-case analysis, and the two remaining genes were positively

associated in our breast cancer-specific case-case analysis: *BRCA1* (OR [95% CI] = 2.38 [1.07, 5.30], $p = 0.0340$) and *BRCA2* (OR [95% CI] = 1.97 [1.22, 3.18], $p = 5.50 \times 10^{-3}$).

3.5 Discussion

We investigated the genetic basis of carcinogenic pleiotropy through whole exome sequencing of individuals diagnosed with multiple primary cancers from two large, multi-ancestry study populations. Comparing individuals with multiple cancers to cancer-free controls uncovered 22 independent, suggestively associated variants, ten of which remained associated when comparing individuals with multiple cancers to those with a single cancer. Across our multiple cancer phenotypes, we also recapitulated previously known gene-based associations in *ATM*, *BRCA1/2*, *CHEK2* and found potentially novel associations in *SAMHD1* and *SLC6A2*. These genes remained associated with multiple cancer diagnoses when comparing to individuals with a single cancer. These findings offer insights into germline exome variants that increase an individual's risk of developing multiple primary cancers.

Compelling findings from our analyses of all individuals with more than one cancer diagnosis include associations with the rare variant rs146381257 in *ZNF106*. Carriers of the rs146381257 risk allele (C) were primarily over-represented in individuals with at least one prostate, breast, lung, or urinary bladder cancer and in individuals with lymphoid neoplasms. Carriers also demonstrated an increased risk of developing multiple cancers compared to individuals with a single cancer. *ZNF106* is an RNA binding protein involved in post-transcriptional regulation and insulin receptor signaling. Although germline variation in *ZNF106* has not previously been associated with cancer risk, a recent study found it to be associated with worse urinary bladder cancer survival³¹.

Additional noteworthy findings from our analyses of all multiple primary cancers combined include cancer susceptibility signals in *SAMHD1* and *SLC6A2*, both having a significantly higher risk being diagnosed with multiple cancers compared to single cancers. Germline *SAMHD1* mutations are implicated in Aicardi-

Goutieres Syndrome (AGS)³², an autosomal recessive condition that results in autoimmune inflammatory encephalopathy. Most cancer-related studies have focused on the role of somatic alternations in *SAMHD1*³³; however, a study of chronic lymphoid leukemia (CLL) proposed an oncogenic role of germline *SAMHD1* variation mediated by DNA repair mechanisms³⁴. Consistent with this hypothesis, we also found increased *SAMHD1* variation in individuals with lymphoid neoplasms, as well as with prostate, breast, colorectal and lung cancers. *SLC6A2*, also known as *NATI*, has been found to be prognostic for colon cancer³⁵, and both in-vivo and in-vitro studies have linked expression to survival in many cancer types, including prostate³⁶ and breast³⁷. Polymorphisms in *SLC6A2* may also interact with smoking exposure to modulate risk for tobacco-related cancers³⁸.

Because we compared multiple primary cancers with both cancer-free controls and individuals diagnosed with a single cancer, we were well positioned to explore patterns of pleiotropy and disentangle variation likely to be driven by single cancers. For example, we identified two variants, rs7872034 (missense variant in *SMC2*) and rs143745791 (missense variant in *NCBPI*), suggestively associated with a diagnosis of at least one breast cancer (plus any other cancer) versus no cancer. These variants remained associated with a diagnosis of breast and another cancer when comparing to individuals diagnosed with a single breast cancer. While rs7872034 is in high LD ($r^2 = 0.98$) with a known breast cancer risk variant (rs4742903; *SMC2* intron)³⁹, it may also increase the risk of developing multiple cancers. Regarding rs143745791, germline variants in *NCBPI* have not been previously associated with cancer; because it is rare (MAF < 0.2%), larger sequencing efforts may be necessary identify variation in studies of individuals with a single cancer. Expression of this gene has been found to promote lung cancer growth and poor prognosis⁴⁰, and *NCBPI* is overexpressed in basal-like and triple-negative breast cancers⁴¹. Similarly, *BRCA1/2* germline variants are prevalent among these subtypes; however, in our study populations, *BRCA1/2* carriers were more common among those with an additional ovarian cancer whereas *NCBPI* carriers more frequently had an additional cervical cancer.

In our prostate cancer-specific analysis comparing individuals with multiple cancers versus those with only a single cancer, we discovered a suggestive association with rs3020779, an eQTL for *RNF123* (also known as *KPCI*), which is a gene involved in p50 mediation and downstream stimulation of multiple tumor suppressors⁴². In our analysis of head and neck cancer, we detected an association with rs12253181 (eQTL for *RTKN2*); while this gene has not previously been associated with head and neck cancer risk, it has been shown to function as an oncogene in non-small cell lung cancer (NSCLC) and decreasing its expression may inhibit proliferation by inducing apoptosis⁴³.

Limitations of our study included the identification of variants that were likely-somatic in our analyses of hematologic cancers due to an expansion of hematopoietic clonal populations with the same acquired mutation (i.e., CHIP). Confounding of germline testing by CHIP has been reported in *TP53*⁴⁴ and *TET2*⁴⁵, so careful interpretation is critical to avoid unnecessary clinical intervention. An additional limitation of our, and other, studies are obtaining accurate effects estimates for rare variants and the reliance on available annotations for inclusion into gene-based tests. Replication of rare findings in larger cohorts and optimization of functional impact annotations could lead to more precise results. Also, while our approach did not allow for formal replication, it was designed to identify signals for a largely understudied phenotype that were concordant in two populations. Finally, while all individuals with multiple cancers were included in our study regardless of genetic ancestry, non-European ancestries were underrepresented; larger, more diverse cohorts will be needed to fully explore the genetic basis of multiple cancers.

Strengths of this work include studying individuals of multiple ancestries who were largely unselected for specific cancer phenotypes. We also performed the first ever exome-wide study of genetic susceptibility to multiple primary cancers, using two large multi-ancestry study populations. Our study design allowed us to characterize variation across multiple primary cancers representing 36 unique sites, as well as to conduct cancer-specific analyses of 16 sites. Using this approach, we confirmed many known single-variant and

gene-based findings, strengthening and supporting our novel results reported for individual cancers through our cancer-specific analyses.

In summary, by undertaking an exome-wide survey of common and rare variation in two large study populations, we identified several variant and gene-based associations that may increase the risk of developing multiple cancers within individuals. Our findings have potential implications for improving our understanding of the shared mechanisms of carcinogenesis. They may also enable screening strategies that prioritize individuals at risk for developing additional cancers. Furthermore, since many of the genes reported here have been considered as potential therapeutic targets in cancer, our work supports the use of germline information to help guide precision medicine. Future studies should aim to replicate our findings and undertake experiments that validate the functionality of the discovered pleiotropic variants. Combined with future research, our results have potential to inform genetic counseling, improve risk prediction for multiple cancers, and guide novel treatment and drug development.

3.6 Acknowledgements

This material is based upon work supported by NIH grant R01 CA201358, RC2 AG036607, and the National Science Foundation Graduate Research Fellowship Program under Grant No. 1650113. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. Support for study enrollment, survey administration, and biospecimen collection of Kaiser Permanente Research Bank participants was provided by the Robert Wood Johnson Foundation, the Wayne and Gladys Valley Foundation, the Ellison Medical Foundation, and Kaiser Permanente national and regional community benefit programs. Additionally, LK is supported by funding from the NCI (K99CA246076) and REG is supported by a Young Investigator Award from the Prostate Cancer Foundation. This research has been conducted using the UK Biobank Resource under Application Number 14015.

3.7 Tables

Table 3.1 Study populations

Table 3.1 Legend: Characteristics of the Kaiser Permanente Research Bank and UK Biobank study populations by ancestry group. Cases are individuals with multiple primary cancers. Controls are those without any cancer.

Ancestry	Population: Kaiser Permanente Research Bank						Population: UK Biobank					
	Cases			Controls			Cases			Controls		
	N	Mean Age	Female (%)	N	Mean Age	Female (%)	N	Mean Age	Female (%)	N	Mean Age	Female (%)
AFR	99	70.5	33.3	100	70.4	32.0	29	55.9	51.7	3,292	51.8	60.4
EAS	95	69.7	49.5	91	69.5	49.5	10	58.8	80.0	1,009	52.6	66.9
EUR	2,786	72.8	43.0	2,815	72.9	43.3	3,249	61.9	51.7	154,047	56.6	54.6
LAT	131	69.5	46.6	130	69.5	45.4	5	63.8	80.0	334	51.8	62.6
SAS	-	-	-	-	-	-	25	58.2	60.0	4,035	53.3	47.0

3.8 Figures

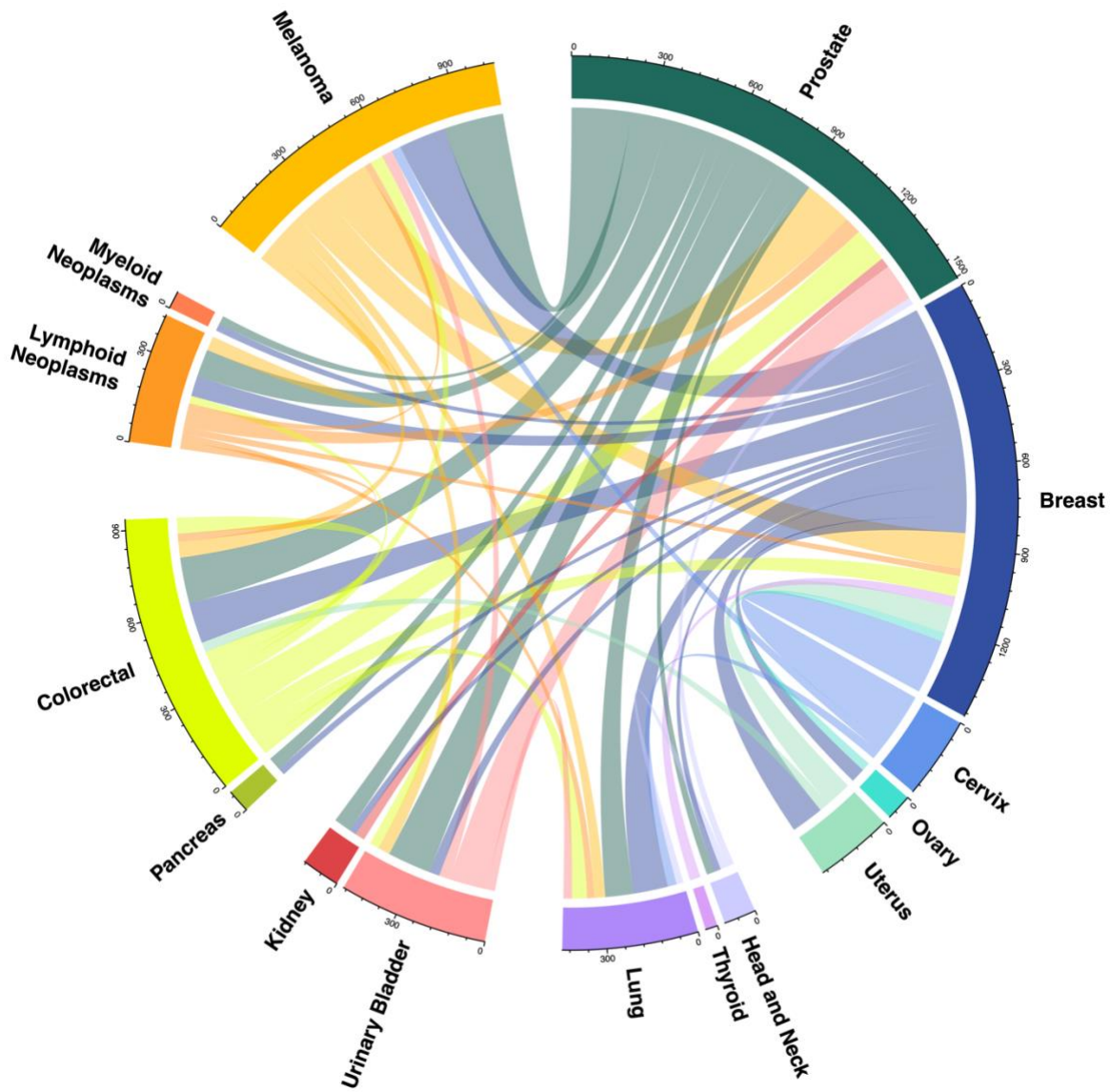


Figure 3.1 Cancer diagnosis pairs present in the combined study populations

Figure 3.1 Legend: Circos plot describing the pairs of first and second cancer diagnoses with at least 25 cases present in Kaiser Permanente Research Bank and the UK Biobank study populations combined. Each connection reflects the number of cases with both of the linked primary cancers, where the color of the line shows the first cancer site diagnosed.

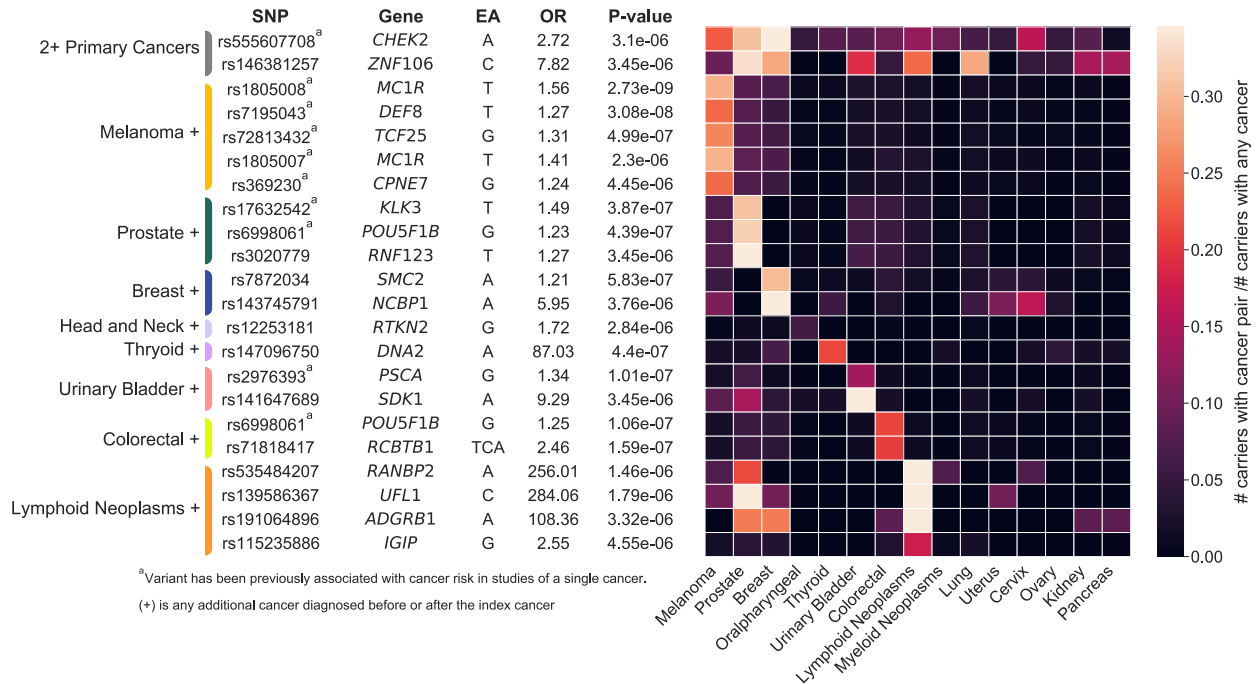


Figure 3.2 Germline single variant association results for multiple primary cancers combined or grouped by organ site

Figure 3.2 Legend: Suggestive ($p < 5 \times 10^{-6}$) germline variant associations with multiple cancer phenotypes versus cancer-free controls ($n = 165,853$) following a fixed-effects meta-analysis of Kaiser Permanente Research Bank and UK Biobank WES data. Associations were detected for any 2+ primary cancers ($n = 6,429$) and with groups of cases defined by a shared index cancer, at any time point, plus any other cancer diagnosis: melanoma + ($n = 1,443$), prostate + ($n = 1,977$), breast + ($n = 1,874$), head and neck + ($n = 283$), thyroid + ($n = 198$), urinary bladder + ($n = 829$), colorectal + ($n = 1,324$), lymphoid neoplasms + ($n = 728$). Variants that have been previously associated in single cancer studies have superscript (a). The heatmap reflects the number of carriers with the risk-increasing allele for each associated variant with the index (y-axis) and additional (x-axis) cancer over the total number of carriers, restricting to cancer cases. When the index and additional cancer are the same, the heatmap value represents all carriers with the specified cancer diagnosis divided by the total number of carriers. Abbreviations: SNP – single nucleotide polymorphism; EA – effect allele; OR – odds ratio.

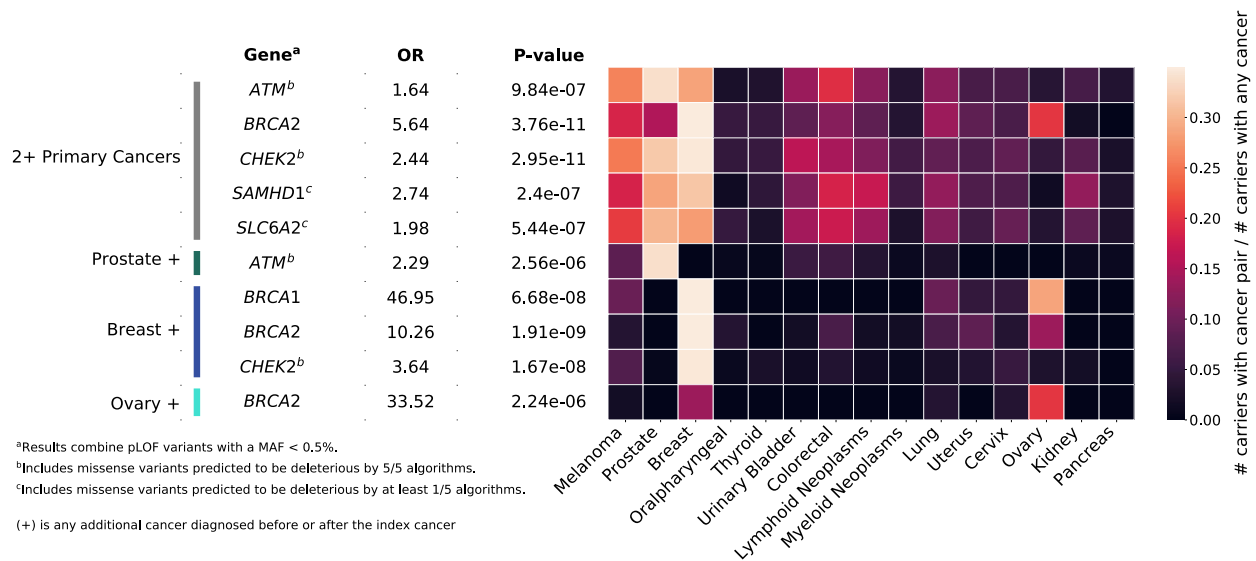


Figure 3.3 Germline gene-based association results for multiple primary cancers combined or grouped by organ site

Figure 3.3 Legend: Burden tests were performed combining variants defined as pLOF with or without deleterious missense variants, defining deleteriousness by at least one (1/5) or all five (5/5) prediction algorithms used (Methods), at a MAF < 0.5%. Following a fixed-effects meta-analysis of Kaiser Permanente Research Bank and UK Biobank data, Bonferroni significant associations ($p < 2.65 \times 10^{-6} = 0.05 / 18,842$) corrected for the number of genes tested were found for comparisons of cancer-free controls ($n = 165,853$) with all cases with any 2+ primary cancers ($n = 6,429$) and with groups of cases defined by an index cancer for the following phenotypes: prostate + ($n = 1,977$), breast + ($n = 1,874$), and ovary + ($n = 239$). For each gene, the variant grouping with the smallest p-value and fewest number of variants was selected. The heatmap reflects the number of carriers of each associated variant, with the index (y-axis) and additional (x-axis) cancer over the total number of carriers, where carrier is defined as having at least one alternate allele across all variants in a given gene, restricting to cancer cases. When the index and additional cancer are the same, the heatmap value represents all carriers with the specified cancer diagnosis divided by the total number of carriers. Abbreviations: OR – odds ratio; pLOF – predicted loss of function.

3.9 Supplementary Materials

Table S3.1 Cancer definitions

Cancer Site	SEER Codes	WHO Codes
Head and Neck	20010, 20011	NA
Esophagus	21010	NA
Stomach	21020	NA
Small Intestine	21030	NA
Colorectal	21040, 21050	NA
Anus	21060	NA
Liver	21070, 21071, 21072	NA
Pancreas	21100	NA
Gallbladder and Biliary Tract	21080, 21090	NA
Other Digestive	21110, 21120, 21130	NA
Lung	22030, 22031	NA
Other Respiratory	22010, 22020, 22050, 22060	NA
Bone	23000	NA
Soft Tissue Sarcoma	24000	NA
Melanoma	25010	NA
Non-Epithelial Skin	25020	NA
Breast	26000	NA
Cervix	27010	NA
Uterus	27020, 27030	NA
Ovary	27040	NA
Other Female Genital	27050, 27060, 27070	NA
Prostate	28010	NA
Testis	28020	NA
Other Male Genital	28030, 28040	NA
Urinary Bladder	29010	NA
Kidney	29020	NA
Other Urinary Organs	29030, 29040	NA
Eye and Orbit	30000	NA
Brain	31010, 31011	NA
Thyroid	32010	NA
Other Endocrine	32020	NA
Mesothelioma	36010	NA
Kaposi Sarcoma	36020	NA
T-Cell and NK-Cell Neoplasms	NA	9729, 9837, 9834, 9831, 9948, 9827, 9719, 9717, 9716, 9708, 9700, 9701, 9718, 9702, 9705, 9714, 9659, 9650, 9665, 9667, 9651, 9652, 9653
Lymphoid Neoplasms	NA	9728, 9836, 9823, 9670, 9833, 9671, 9689, 9940, 9732, 9731, 9699, 9699, 9690, 9691, 9695, 9698, 9673, 9680, 9679, 9678, 9687, 9826
Myeloid Neoplasms	NA	9875, 9963, 9964, 9961, 9950, 9962, 9975, 9945, 9876, 9946, 9980, 9982, 9980, 9985, 9983, 9986, 9989, 9896, 9866, 9871, 9897, 9895, 9895, 9895, 9920, 9920, 9920, 9920, 9861, 9872, 9873, 9874, 9867, 9891, 9840, 9910, 9870, 9931, 9805

Table S3.2 Cancer pairs with at least 25 cases combined across study populations

Cancer Dx 1	Cancer Dx 2	Num Cases UKB	Num Cases Kaiser	Total Cases
Prostate	Melanoma	47	174	221
Cervix	Breast	146	56	202
Melanoma	Prostate	60	120	180
Breast	Melanoma	65	109	174
Prostate	Colorectal	88	82	170
Prostate	Urinary Bladder	46	115	161
Breast	Colorectal	86	65	151
Urinary Bladder	Prostate	86	53	139
Melanoma	Breast	70	62	132
Colorectal	Prostate	65	65	130
Breast	Lung	68	60	128
Breast	Uterus	63	58	121
Prostate	Lung	45	59	104
Prostate	Lymphoid Neoplasms	34	68	102
Uterus	Breast	50	48	98
Colorectal	Breast	39	37	76
Breast	Lymphoid Neoplasms	38	37	75
Prostate	Kidney	27	37	64
Lymphoid Neoplasms	Prostate	32	31	63
Colorectal	Colorectal	46	13	59
Melanoma	Colorectal	29	30	59
Colorectal	Lung	32	23	55
Breast	Ovary	25	29	54
Melanoma	Lymphoid Neoplasms	24	29	53
Prostate	Pancreas	18	32	50
Colorectal	Melanoma	19	29	48
Melanoma	Urinary Bladder	8	37	45
Melanoma	Lung	19	22	41
Thyroid	Breast	19	21	40
Urinary Bladder	Melanoma	16	24	40
Breast	Urinary Bladder	21	17	38
Colorectal	Urinary Bladder	17	19	36
Kidney	Prostate	21	15	36
Uterus	Colorectal	14	22	36
Cervix	Melanoma	26	9	35
Cervix	Lung	27	6	33
Ovary	Breast	17	15	32
Urinary Bladder	Lung	13	19	32
Prostate	Myeloid Neoplasms	4	27	31
Lymphoid Neoplasms	Colorectal	20	9	29
Prostate	Head and Neck	7	22	29
Colorectal	Lymphoid Neoplasms	13	15	28
Breast	Myeloid Neoplasms	15	12	27
Lymphoid Neoplasms	Breast	16	11	27
Breast	Kidney	15	11	26
Breast	Pancreas	15	11	26
Lymphoid Neoplasms	Lung	17	9	26
Head and Neck	Lung	15	11	26
Head and Neck	Prostate	11	15	26
Breast	Head and Neck	19	6	25
Lymphoid Neoplasms	Melanoma	11	14	25

Table S3.3 Groupings of multiple cancer cases by a shared index cancer

Index Cancer	Num Cases UKB	Num Cases KPRB	Total Cases
Prostate	809	1168	1977
Breast	1030	844	1874
Melanoma	547	896	1443
Colorectal	745	579	1324
Urinary Bladder	347	482	829
Lung	401	389	790
Lymphoid Neoplasms	366	362	728
Cervix	408	127	535
Uterus	214	243	457
Kidney	229	189	418
Head and Neck	145	138	283
Myeloid Neoplasms	111	142	253
Ovary	142	97	239
Pancreas	93	105	198
Thyroid	91	107	198
Esophagus	99	39	138
Stomach	88	42	130
Other Female Genital	63	57	120
Brain	71	40	111
T-Cell and NK-Cell Neoplasms	70	37	107
Eye and Orbit	89	16	105
Other Respiratory	55	44	99
Soft Tissue Sarcoma	51	47	98
Other Urinary Organs	38	39	77
Non-Epithelial Skin	25	51	76
Anus	33	38	71
Liver	42	29	71
Testis	59	10	69
Small Intestine	36	25	61
Gallbladder and Biliary Tract	30	19	49
Mesothelioma	30	11	41
Other Digestive	19	15	34
Other Male Genital	14	16	30
Other Endocrine	15	9	24
Bone	11	7	18
Kaposi Sarcoma	1	3	4

Table S3.4 Single-variant case-control meta-analysis associations for multiple cancer phenotypes

Phenotype	SNP	Chrom	Position	Gene	EA	Beta	SE	Pvalue	Cochran's Q	Pvalue Q
Multiple Cancers	rs555607708	22	28695868	CHEK2	A	1.0	0.22	3.1E-06	3.0	0.082
Multiple Cancers	rs146381257	15	42448087	ZNF106	C	2.1	0.44	3.5E-06	0.62	0.43
Melanoma	rs1805008	16	89919736	MC1R	T	0.45	0.075	2.7E-09	0.69	0.41
Melanoma	rs7195043	16	89954453	DEF8	T	0.24	0.043	3.1E-08	0.015	0.90
Melanoma	rs72813432	16	89898544	TCF25	G	0.27	0.054	5.0E-07	0.11	0.75
Melanoma	rs1805007	16	89919709	MC1R	T	0.34	0.072	2.3E-06	1.8	0.18
Melanoma	rs369230	16	89579029	CPNE7	G	0.21	0.046	4.5E-06	2.4	0.12
Prostate	rs17632542	19	50858501	KLK3	T	0.40	0.078	3.9E-07	0.77	0.38
Prostate	rs6998061	8	127416393	POU5F1B	G	0.20	0.040	4.4E-07	0.11	0.74
Prostate	rs3020779	3	49687375	RNF123	T	0.24	0.051	3.5E-06	0.035	0.85
Breast	rs7872034	9	104134528	SMC2	A	0.19	0.038	5.8E-07	2.5	0.11
Breast	rs143745791	9	97643352	NCBP1	A	1.8	0.39	3.8E-06	1.1	0.30
Head and Neck	rs12253181	10	62195767	RTKN2	G	0.54	0.12	2.8E-06	0.21	0.65
Thyroid	rs147096750	10	68430491	DNA2	A	4.5	0.88	4.4E-07	0.57	0.45
Urinary Bladder	rs2976393	8	142682200	PSCA	G	0.29	0.055	1.0E-07	4.1E-06	1.0
Urinary Bladder	rs141647689	7	3962611	SDK1	A	2.2	0.48	3.5E-06	0.20	0.66
Colorectal	rs6998061	8	127416393	POU5F1B	G	0.23	0.043	1.1E-07	0.19	0.67
Colorectal	rs71818417	13	49546318	RCBTB1	TCA	0.90	0.17	1.6E-07	0.18	0.67
Lymphoid Neoplasms	rs535484207	2	108720130	RANBP2	A	5.6	1.2	1.5E-06	1.1	0.30
Lymphoid Neoplasms	rs139586367	6	96526408	UFL1	C	5.7	1.2	1.8E-06	0.54	0.46
Lymphoid Neoplasms	rs191064896	8	142520791	ADGRB1	A	4.7	1.0	3.3E-06	0.0070	0.93
Lymphoid Neoplasms	rs115235886	5	140128528	IGIP	G	0.94	0.20	4.6E-06	0.070	0.79
Lymphoid Neoplasms	rs188201162	1	148979731	PDE4DIP	T	1.1	0.25	4.6E-06	0.032	0.86
Myeloid Neoplasms	rs756958159	20	32434638	ASXL1	AG	5.2	0.77	1.5E-11	0.61	0.43
Myeloid Neoplasms	rs751713049	17	76736877	SRSF2	T	7.0	1.1	6.7E-10	2.3	0.13
Myeloid Neoplasms	rs559063155	2	197402110	SF3B1	C	7.7	1.6	2.4E-06	0.071	0.79

Table S3.5 Gene-based case-control meta-analysis associations for multiple cancer phenotypes

Phenotype	Gene	Mask	Beta	SE	Pvalue	Cochran's Q	Pvalue-Q
Multiple Cancers	SLC6A2	pLOF+missense (1/5)	0.69	0.14	5.4E-07	1.9	0.17
Multiple Cancers	ATM	pLOF+missense (5/5)	0.49	0.10	9.8E-07	2.1	0.15
Multiple Cancers	CHEK2	pLOF+missense (5/5)	0.89	0.13	3.0E-11	3.3	0.068
Multiple Cancers	SAMHD1	pLOF+missense (1/5)	1.0	0.20	2.4E-07	5.4	0.021
Multiple Cancers	BRCA2	pLOF	1.7	0.26	3.8E-11	22	3.3E-06
Prostate	ATM	pLOF+missense (5/5)	0.83	0.18	2.6E-06	0.52	0.47
Breast	CHEK2	pLOF+missense (5/5)	1.3	0.23	1.7E-08	0.0080	0.93
Breast	BRCA1	pLOF	3.9	0.71	6.7E-08	10	1.5E-03
Breast	BRCA2	pLOF	2.3	0.39	1.9E-09	17	3.6E-05
Ovary	BRCA2	pLOF	3.5	0.74	2.2E-06	9.6	2.0E-03
Myeloid Neoplasms	ASXL1	pLOF	4.5	0.62	3.5E-13	0.25	0.62
Myeloid Neoplasms	TET2	pLOF	3.4	0.54	5.5E-10	0.58	0.45
Myeloid Neoplasms	JAK2	pLOF+missense (5/5)	3.7	0.50	3.7E-13	0.94	0.33
Myeloid Neoplasms	DDX41	pLOF+missense (5/5)	3.8	0.78	1.0E-06	1.2	0.28

Table S3.6 Case-case analysis for significant single-variant associations

Phenotype	SNP	Chrom	Position	Gene	EA	Beta	SE	Pvalue
Multiple Cancers	rs555607708	22	28695868	CHEK2	A	0.45	0.18	0.015
Multiple Cancers	rs146381257	15	42448087	ZNF106	C	1.7	0.83	0.042
Melanoma	rs1805008	16	89919736	MC1R	T	0.15	0.10	0.13
Melanoma	rs7195043	16	89954453	DEF8	T	0.010	0.069	0.89
Melanoma	rs72813432	16	89898544	TCF25	G	0.15	0.080	0.054
Melanoma	rs1805007	16	89919709	MC1R	T	0.015	0.095	0.88
Melanoma	rs369230	16	89579029	CPNE7	G	0.095	0.071	0.18
Prostate	rs17632542	19	50858501	KLK3	T	-0.10	0.12	0.41
Prostate	rs6998061	8	127416393	POU5F1B	G	0.048	0.057	0.40
Prostate	rs3020779	3	49687375	RNF123	T	0.14	0.066	0.038
Breast	rs7872034	9	104134528	SMC2	A	0.15	0.048	2.5E-03
Breast	rs143745791	9	97643352	NCBP1	A	1.3	0.29	8.4E-06
Head and Neck	rs12253181	10	62195767	RTKN2	G	0.69	0.17	7.5E-05
Thyroid	rs147096750	10	68430491	DNA2	A	2.3	0.98	0.020
Urinary Bladder	rs2976393	8	142682200	PSCA	G	0.14	0.095	0.13
Urinary Bladder	rs141647689	7	3962611	SDK1	A	1.2	0.48	0.013
Colorectal	rs6998061	8	127416393	POU5F1B	G	0.083	0.061	0.17
Colorectal	rs71818417	13	49546318	RCBTB1	TCA	1.3	0.45	4.5E-03
Lymphoid Neoplasms	rs535484207	2	108720130	RANBP2	A	3.0	1.7	0.073
Lymphoid Neoplasms	rs139586367	6	96526408	UFL1	C	1.9	2.3	0.41
Lymphoid Neoplasms	rs191064896	8	142520791	ADGRB1	A	1.6	1.5	0.29
Lymphoid Neoplasms	rs115235886	5	140128528	IGIP	G	0.77	0.25	2.0E-03
Lymphoid Neoplasms	rs188201162	1	148979731	PDE4DIP	T	NA	NA	NA
Myeloid Neoplasms	rs756958159	20	32434638	ASXL1	AG	0.97	0.77	0.21
Myeloid Neoplasms	rs751713049	17	76736877	SRSF2	T	0.38	0.69	0.59
Myeloid Neoplasms	rs559063155	2	197402110	SF3B1	C	0.37	1.3	0.77

Table S3.7 Case-case analysis for significant gene-based associations

Phenotype	Gene	Mask	Beta	SE	Pvalue
Multiple Cancers	SLC6A2	pLOF+missense (1/5)	0.62	0.13	3.9E-06
Multiple Cancers	ATM	pLOF+missense (5/5)	0.36	0.11	1.1E-03
Multiple Cancers	CHEK2	pLOF+missense (5/5)	0.45	0.12	2.3E-04
Multiple Cancers	SAMHD1	pLOF+missense (1/5)	0.44	0.16	5.3E-03
Multiple Cancers	BRCA2	pLOF	0.62	0.18	5.4E-04
Prostate	ATM	pLOF+missense (5/5)	0.60	0.21	4.6E-03
Breast	CHEK2	pLOF+missense (5/5)	0.23	0.21	0.29
Breast	BRCA1	pLOF	0.87	0.41	0.034
Breast	BRCA2	pLOF	0.68	0.24	5.5E-03
Ovary	BRCA2	pLOF	0.54	0.47	0.25
Myeloid Neoplasms	ASXL1	pLOF	0.51	0.68	0.45
Myeloid Neoplasms	TET2	pLOF	-1.3	1.1	0.23
Myeloid Neoplasms	JAK2	pLOF+missense (5/5)	-0.73	0.48	0.13
Myeloid Neoplasms	DDX41	pLOF+missense (5/5)	0.63	0.55	0.25

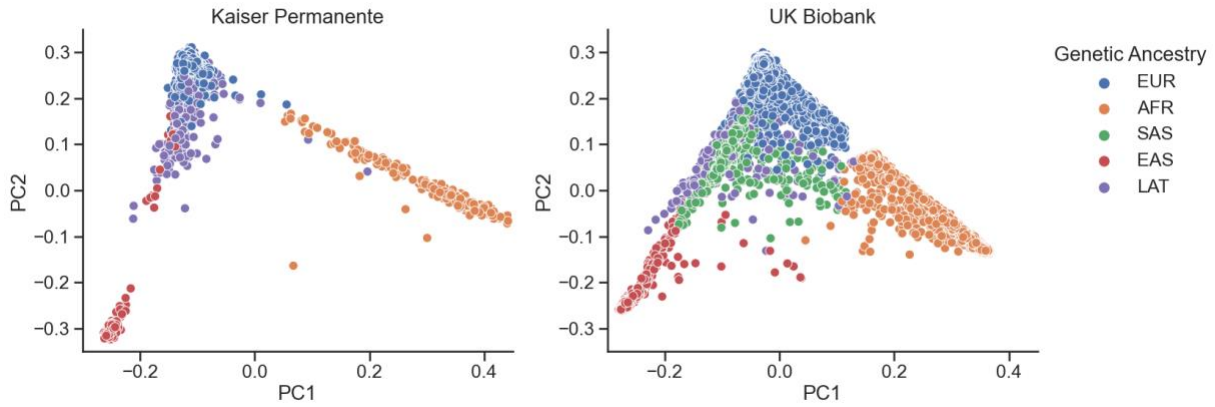


Figure S3.1 Genetic ancestry in the Kaiser Permanente Research Bank and UK Biobank

Figure S3.1 Legend: Principal components (PCs) were computed using imputed and quality-controlled genotype array data with flashPCA2. To enable accurate ancestry assignment, PCs were projected onto the 1000G phase 3 reference populations. Individuals were assigned to genetic ancestry populations using the minimum distance to the top 10 PCs, and outliers with PCs greater than 5 standard deviations from the assigned population mean were removed. Abbreviations: EUR – European; AFR – African; SAS – South Asian; EAS – East Asian; LAT – Hispanic/Latino.

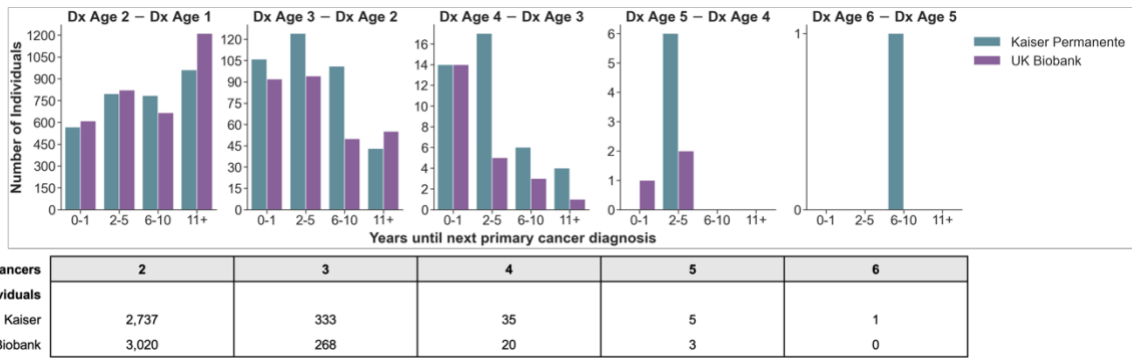
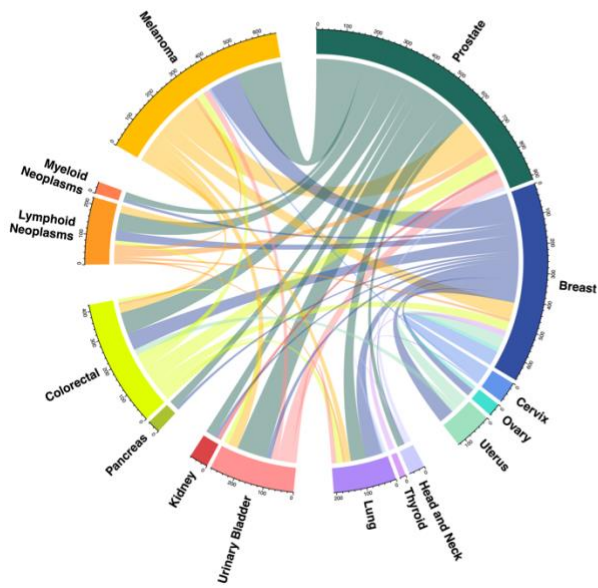


Figure S3.2 Number of primary cancer diagnoses and time intervals between cancer diagnoses for Kaiser Permanente Research Bank and UK Biobank individuals with multiple cancers

Figure S3.2 Legend: The number of primary cancer diagnoses for each individual was tabulated for the Kaiser Permanente Research Bank and UK Biobank study populations. The above bar plot shows individuals binned according to time between cancer diagnoses, reflecting the subsequent diagnosis occurring up to 1 year, 2 to 5 years, 6 to 10 years, or more than 11 years after the prior diagnosis. Abbreviations: Dx – Diagnosis

Population: Kaiser Permanente



Population: UK Biobank

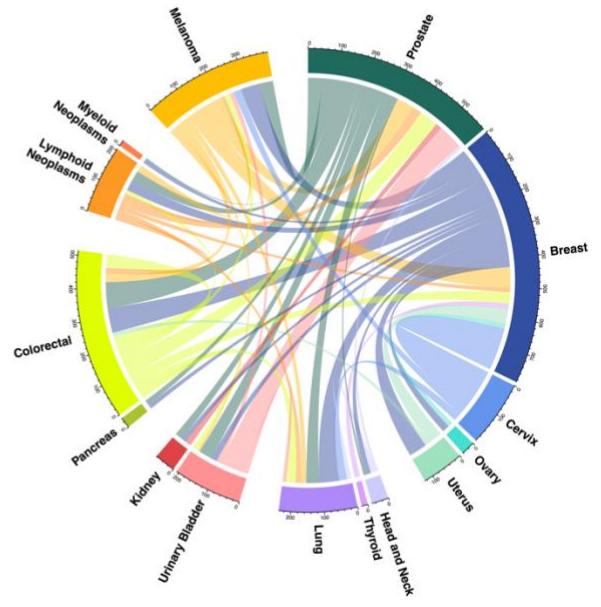


Figure S3.3 Most common cancer pairs present in Kaiser Permanente and UK Biobank cases with multiple cancers

Figure S3.3 Legend: Circos plots describing pairs of first and second cancer diagnoses with at least 25 cases present in the Kaiser Permanente Research Bank (left) and UK Biobank (right). Each connection reflects the number of individuals with both of the linked primary cancers, where the color of the line shows the first cancer site diagnosed.

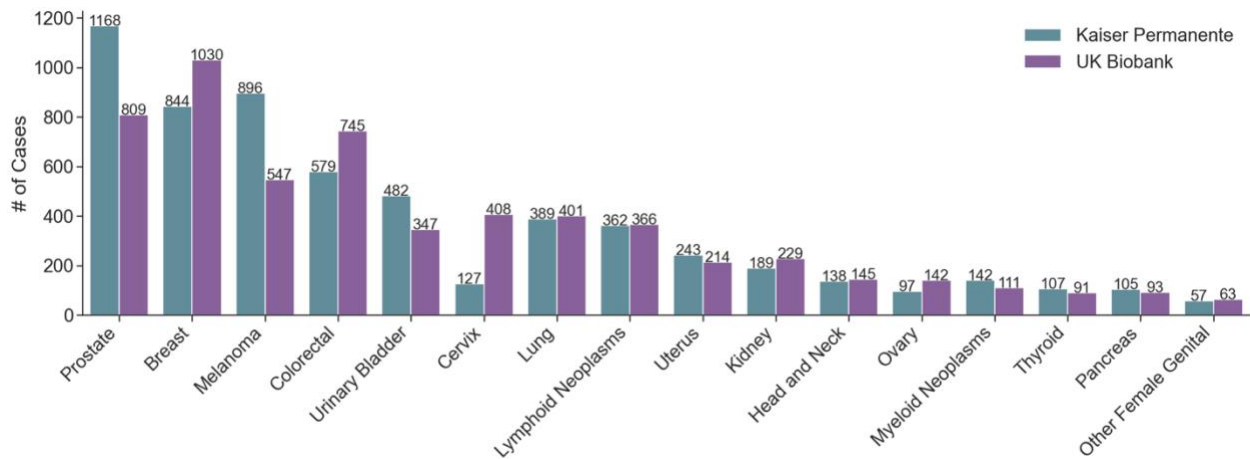


Figure S3.4 Cancers represented in the Kaiser Permanente Research Bank and UK Biobank with sufficient sample size for exome-wide association analyses

Figure S3.4 Legend: All individuals with multiple cancer who were diagnosed with at a specific site, at any time point, were grouped together. A total of 16 cancer sites (represented above) had sufficient sample size ($N > 50$) in each study population for downstream analyses.

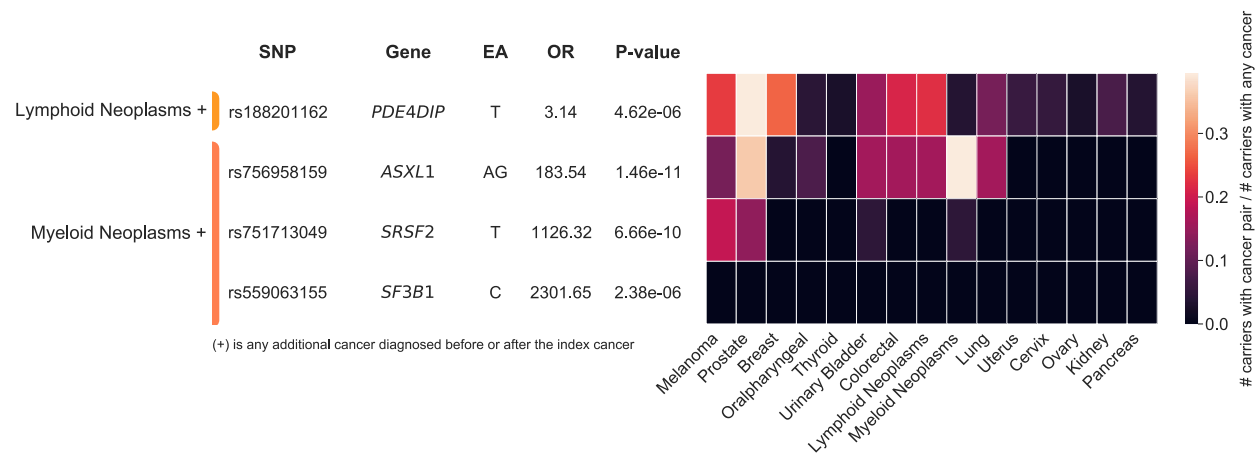


Figure S3.5 Significant single variant association results due to CHIP

Figure S3.5 Legend: Significant variant associations ($p < 5 \times 10^{-6}$) with blood cancer phenotypes compared to cancer-free controls following a fixed-effects meta-analysis of the Kaiser Permanente Research Bank and UK Biobank WES data. Because of low allele balance, which is suggestive of confounding by clonal hematopoiesis of indeterminate potential (CHIP), mutations are expected to be somatic. The heatmap reflects the number of carriers with the risk-increasing allele for each associated variant with the index (y-axis) and additional (x-axis) cancer over the total number of carriers. When the index and additional cancer are the same, the heatmap value represents all carriers with the specified cancer diagnosis divided by the total number of carriers, restricting to cancer cases.

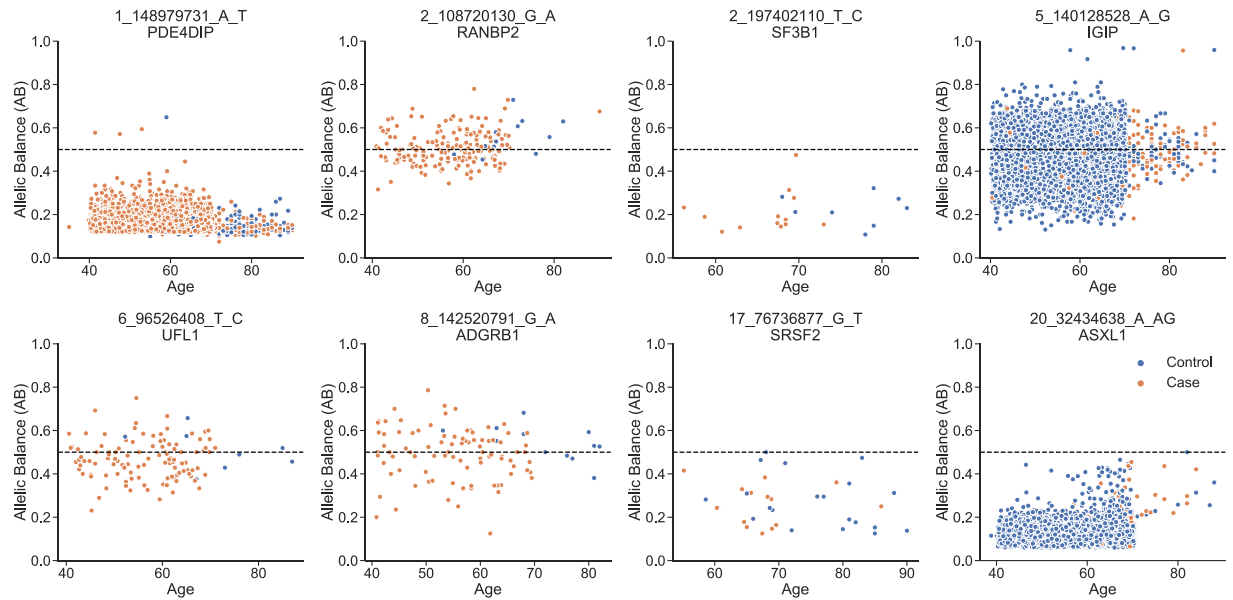


Figure S3.6 Allele balance for findings related to lymphoid and myeloid neoplasms

Figure S6 Legend: For significant variants discovered in our cancer-specific analyses of either myeloid or lymphoid neoplasms, we looked at the allele balance as a function of age for individuals heterozygous at each locus. If the majority of heterozygous individuals had allele balance below 0.5, we assumed that they were likely somatic due to confounding by clonal hematopoiesis of indeterminate potential (CHIP).

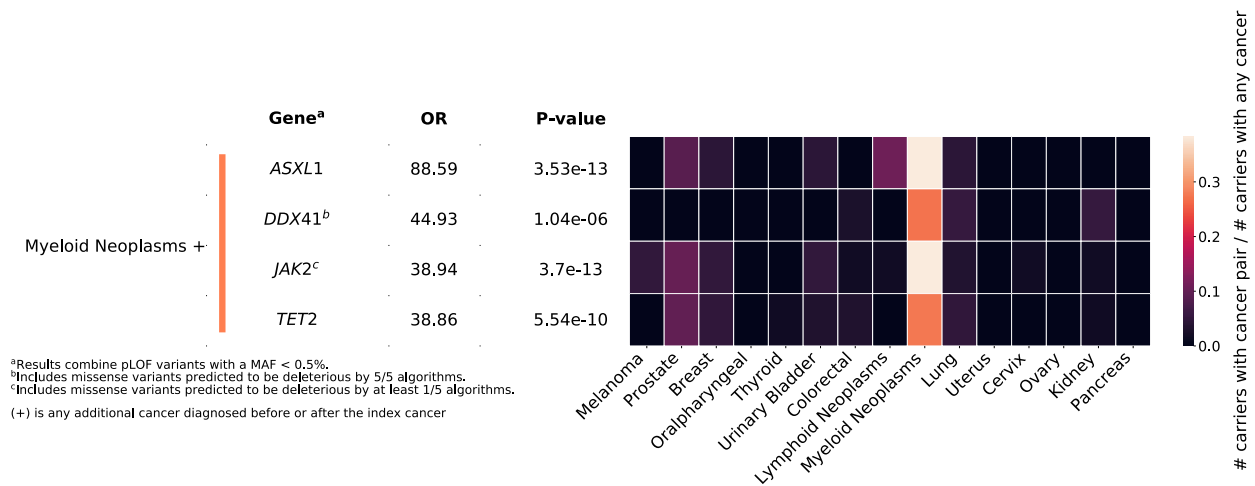


Figure S3.7 Significant gene-based association results due to CHIP

Figure S7 Legend: Burden tests were performed combining variants defined as pLOF with or without deleterious missense variants, defining deleteriousness by at least one (1/5) or all five (5/5) of prediction algorithms used (Methods), at a minor allele frequency < 0.5%. Bonferroni significant associations ($P < 2.65 \times 10^{-6} = 0.05 / 18,842$) corrected for the number of genes tested were found for comparisons of cancer-free controls with myeloid neoplasms following a fixed-effects meta-analysis of Kaiser Permanente Research Bank and UK Biobank data. For each gene, the variant grouping with the smallest p-value and fewest number of variants was selected. The heatmap reflects the number of carriers of each associated variant, with the index (y-axis) and additional (x-axis) cancer over the total number of carriers where carrier is defined as having at least one alternate allele across all variants in a given gene. When the index and additional cancer are the same, the heatmap value represents all carriers with the specified cancer diagnosis divided by the total number of carriers, restricting to cancer cases. Myeloid associations occur in frequently mutated clonal hematopoiesis of indeterminate potential (CHIP) genes.

References

1. Vogt, A. *et al.* Multiple primary tumours: challenges and approaches, a review. *ESMO Open* **2**, e000172 (2017).
2. Copur, M. S. & Manapuram, S. Multiple Primary Tumors Over a Lifetime. *Oncology (Williston Park)* **33**, 629384 (2019).
3. Gaspar, T. B. *et al.* Telomere Maintenance Mechanisms in Cancer. *Genes* **9**, 241 (2018).
4. Smedby, K. E. *et al.* GWAS of Follicular Lymphoma Reveals Allelic Heterogeneity at 6p21.32 and Suggests Shared Genetic Susceptibility with Diffuse Large B-cell Lymphoma. *PLoS Genet* **7**, e1001378 (2011).
5. Karnes, J. H. *et al.* Phenome-wide scanning identifies multiple diseases and disease severity phenotypes associated with HLA variants. *Sci. Transl. Med.* **9**, eaai8708 (2017).
6. Huppi, K., Pitt, J. J., Wahlberg, B. M. & Caplen, N. J. The 8q24 Gene Desert: An Oasis of Non-Coding Transcriptional Activity. *Front. Gene.* **3**, (2012).
7. Rashkin, S. R. *et al.* Pan-cancer study detects genetic risk variants and shared genetic basis in two large cohorts. *Nat Commun* **11**, 4423 (2020).
8. Lindström, S. *et al.* Quantifying the Genetic Correlation between Multiple Cancer Types. *Cancer Epidemiol Biomarkers Prev* **26**, 1427–1435 (2017).
9. Hoffmann, T. J. *et al.* Imputation of the Rare HOXB13 G84E Mutation and Cancer Risk in a Large Population-Based Cohort. *PLoS Genet* **11**, e1004930 (2015).
10. Szustakowski, J. D. *et al.* Advancing human genetics research and drug discovery through exome sequencing of the UK Biobank. *Nat Genet* **53**, 942–948 (2021).
11. Graff, R. E. *et al.* Cross-cancer evaluation of polygenic risk scores for 16 cancer types in two large cohorts. *Nat Commun* **12**, 970 (2021).
12. Adamo, M., Groves, C., Dickie, L. & Ruhl, J. SEER Program Coding and Staging Manual 2021. *National Cancer Institute, Bethesda, MD 20892.* (2020).

13. Harris, N. L. *et al.* The World Health Organization Classification of Neoplasms of the Hematopoietic and Lymphoid Tissues: Report of the Clinical Advisory Committee Meeting – Airlie House, Virginia, November, 1997. *Hematol J* **1**, 53–66 (2000).
14. Abraham, G., Qiu, Y. & Inouye, M. FlashPCA2: principal component analysis of Biobank-scale genotype datasets. *Bioinformatics* **33**, 2776–2778 (2017).
15. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
16. Geisinger-Regeneron DiscovEHR Collaboration *et al.* Exome sequencing and characterization of 49,960 individuals in the UK Biobank. *Nature* **586**, 749–756 (2020).
17. Yun, T. *et al.* Accurate, scalable cohort variant calls using DeepVariant and GLnexus. *Bioinformatics* **36**, 5582–5589 (2021).
18. Mbatchou, J. *et al.* Computationally efficient whole-genome regression for quantitative and binary traits. *Nat Genet* **53**, 1097–1103 (2021).
19. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* **6**, 80–92 (2012).
20. Dong, C. *et al.* Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Human Molecular Genetics* **24**, 2125–2137 (2015).
21. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research* **38**, e164–e164 (2010).
22. Backman, J. D. *et al.* Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature* **599**, 628–634 (2021).
23. Han, B. & Eskin, E. Random-Effects Model Aimed at Discovering Associations in Meta-Analysis of Genome-wide Association Studies. *The American Journal of Human Genetics* **88**, 586–598 (2011).
24. Viechtbauer, W. Conducting Meta-Analyses in R with the **metafor** Package. *J. Stat. Soft.* **36**, (2010).

25. Steensma, D. P. *et al.* Clonal hematopoiesis of indeterminate potential and its distinction from myelodysplastic syndromes. *Blood* **126**, 9–16 (2015).
26. Cybulski, C. *et al.* CHEK2 Is a Multiorgan Cancer Susceptibility Gene. *The American Journal of Human Genetics* **75**, 1131–1135 (2004).
27. Amos, C. I. *et al.* Genome-wide association study identifies novel loci predisposing to cutaneous melanoma. *Hum Mol Genet* **20**, 5012–5023 (2011).
28. Li, H., Fei, X., Shen, Y. & Wu, Z. Association of gene polymorphisms of KLK3 and prostate cancer: A meta-analysis. *Adv Clin Exp Med* **29**, 1001–1009 (2020).
29. Hazelett, D. J. *et al.* Comprehensive functional annotation of 77 prostate cancer risk loci. *PLoS Genet* **10**, e1004102 (2014).
30. Hutter, C. M. *et al.* Characterization of the association between 8q24 and colon cancer: gene-environment exploration and meta-analysis. *BMC Cancer* **10**, 670 (2010).
31. Wu, Y. *et al.* Identification of the Functions and Prognostic Values of RNA Binding Proteins in Bladder Cancer. *Front. Genet.* **12**, 574196 (2021).
32. Martinez-Lopez, A. *et al.* SAMHD1 deficient human monocytes autonomously trigger type I interferon. *Molecular Immunology* **101**, 450–460 (2018).
33. Mauney, C. H. & Hollis, T. SAMHD1: Recurring roles in cell cycle, viral restriction, cancer, and innate immunity. *Autoimmunity* **51**, 96–110 (2018).
34. Clifford, R. *et al.* SAMHD1 is mutated recurrently in chronic lymphocytic leukemia and is involved in response to DNA damage. *Blood* **123**, 1021–1031 (2014).
35. Shi, C. *et al.* Hypermethylation of N-Acetyltransferase 1 Is a Prognostic Biomarker in Colon Adenocarcinoma. *Front. Genet.* **10**, 1097 (2019).
36. Tiang, J. M., Butcher, N. J., Cullinane, C., Humbert, P. O. & Minchin, R. F. RNAi-Mediated Knock-Down of Arylamine N-acetyltransferase-1 Expression Induces E-cadherin Up-Regulation and Cell-Cell Contact Growth Inhibition. *PLoS ONE* **6**, e17031 (2011).

37. Minchin, R. F. & Butcher, N. J. Trimodal distribution of arylamine N-acetyltransferase 1 mRNA in breast cancer tumors: association with overall survival and drug resistance. *BMC Genomics* **19**, 513 (2018).
38. McKay, J. D. *et al.* Sequence Variants of NAT1 and NAT2 and Other Xenometabolic Genes and Risk of Lung and Aerodigestive Tract Cancers in Central Europe. *Cancer Epidemiology Biomarkers & Prevention* **17**, 141–147 (2008).
39. kConFab Investigators *et al.* Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses. *Nat Genet* **52**, 572–581 (2020).
40. Zhang, H. *et al.* NCBP1 promotes the development of lung adenocarcinoma through up-regulation of CUL4B. *J Cell Mol Med* **23**, 6965–6977 (2019).
41. Wang, L. *et al.* Novel RNA-Affinity Proteogenomics Dissects Tumor Heterogeneity for Revealing Personalized Markers in Precision Prognosis of Cancer. *Cell Chemical Biology* **25**, 619-633.e5 (2018).
42. Kravtsova-Ivantsiv, Y. *et al.* Excess of the NF- κ B p50 subunit generated by the ubiquitin ligase KPC1 suppresses tumors via PD-L1– and chemokines-mediated mechanisms. *Proc Natl Acad Sci USA* **117**, 29823–29831 (2020).
43. Ji, L. *et al.* RTKN2 is Associated with Unfavorable Prognosis and Promotes Progression in Non-Small-Cell Lung Cancer. *OTT Volume* **13**, 10729–10738 (2020).
44. Weitzel, J. N. *et al.* Somatic TP53 variants frequently confound germ-line testing results. *Genetics in Medicine* **20**, 809–816 (2018).
45. Tulstrup, M. *et al.* TET2 mutations are associated with hypermethylation at key regulatory enhancers in normal and malignant hematopoiesis. *Nat Commun* **12**, 6061 (2021).

Funding and Support

This work was supported by the National Science Foundation Graduate Research Fellowship program under grant No. 1650113, NIH grant No. CA201358, RC2 AG036607, and the UCSF Discovery Fellows Program.

Publishing Agreement

It is the policy of the University to encourage open access and broad distribution of all theses, dissertations, and manuscripts. The Graduate Division will facilitate the distribution of UCSF theses, dissertations, and manuscripts to the UCSF Library for open access and distribution. UCSF will make such theses, dissertations, and manuscripts accessible to the public and will take reasonable steps to preserve these works in perpetuity.

I hereby grant the non-exclusive, perpetual right to The Regents of the University of California to reproduce, publicly display, distribute, preserve, and publish copies of my thesis, dissertation, or manuscript in any form or media, now existing or later derived, including access online for teaching, research, and public service purposes.

DocuSigned by:

Taylor Canazos

A96D18F4A7194E5...

Author Signature

4/4/2022

Date