UNIVERSITY OF CALIFORNIA SAN DIEGO

Using Germline Variation to Study Inter-Individual Variability in Cancer Risk and Host Anti-Tumor Immune Response

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

in

Biomedical Sciences

by

Meghana Sai Pagadala

Committee in charge:

Professor Hannah Carter, Chair
Professor Jill Mesirov, Co-Chair
Professor Silvio Gutkind
Professor Jason Sicklick
Professor Pandurangan Vijayanand
Professor Robert Wechsler-Reya

2022

The dissertation of Meghana Sai Pagadala is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2022

DEDICATION

I dedicate this dissertation to my family, friends and mentors. I would not be here if not for their tremendous support and love.

I also dedicate this to my grandfather who would have been so proud of me.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank the incredible mentors and colleagues that made this work possible. First, I must acknowledge the tremendous support Dr. Hannah Carter has given me over the past 4 years. I asked to join her lab with minimal computational experience, yet she took a chance on me and helped me develop critical skills, such as creativity, curiosity and independence needed to be a successful researcher. I want to also thank Dr. Jill Mesirov for her input and advice. Dr. Mesirov pushed me to be detail-oriented and thorough, strengthening my skills as a researcher. Both Drs. Carter and Mesirov are incredible role models not just in science, but also in life, and I cannot thank them enough. I hope that we continue to work together in the future.

I would like to thank the rest of my committee, Drs. Gutkind, Pandurangan, Wechsler-Reya and Sicklick for their guidance and feedback. During my graduate education, I received the opportunity to work with other groups that truly made my graduate experience special. I would like to thank Drs. Wechsler-Reya, Mesirov, Carter for inviting me to the Rady Children's Hospital Molecular Tumor Board. Working on the Rady Children's Hospital Molecular Tumor Board was an amazing experience as a future physician-scientist. Also, I must thank the members of the Million Veteran Program 022 working group, Drs. Hauger, Lynch, Panizzon, Rose, Seibert and Kaur, for being incredibly supportive. Not only did they offer an opportunity to analyze one of the largest and most diverse genetic cohorts, but they also made every Wednesday MVP meeting enjoyable and inspiring. I must thank Dr. Wang and the rest of the Cancer Biology, Informatics and Omics (CBIO) training grant mentors for their guidance during the early part of my graduate education. Dr. Wang and her support helped me pivot to bioinformatics and her feedback during CBIO presentations was greatly helpful. I would like to thank the Cancer Genomics Journal Club advisors, Drs. Harismendy, Curtius, Carter and Mesirov, for great discussion every Tuesday. I

would also like to acknowledge the leaders and administrators of the UC San Diego Medical Scientist Training Program, the UC San Diego Biomedical Sciences Graduate Program, and the UC San Diego School of Medicine. It truly took a village to get to this point.

I thank the National Cancer Institute for their generous grant support through the CBIO training grant and F30.

Lastly, I would like to thank the members of the Carter and Mesirov labs, who made my graduate experience so joyful, fun, and enriching. My friends and family were critical during this whole process, and I cannot thank them enough.

The content of Chapter 1, in full, is a reprint of the material as it appears in biorxiv. Pagadala, Meghana; Wu, Victoria; Pérez-Guijarro, Eva; Kim, Hyo; Castro, Andrea; Talwar, James; Gonzalez-Colin, Cristian; Cao, Steven; Schmiedel, Benjamin; Salem, Rany; Morris, Gerald; Harismendy, Olivier; Patel, Sandip; Mesirov, Jill; Zanetti, Maurizio; Day, Chi-Ping; Fan, Chun; Thompson, Wesley; Merlino, Glenn; Gutkind, Silvio; Vijayanand, Panduranga; Carter, Hannah. The dissertation author was the primary investigator and author of this paper.

The content of Chapter 2, in part, is currently being prepared for submission for publication of the material. Pagadala; Meghana; Castro, Andrea; Sears, Timothy; Chavez, Lukas; Chapman, Owen; Wechsler-Reya, Robert; Mesirov, Jill*; Carter, Hannah*. The dissertation author was the primary author of this material.

The content of Chapter 3, in full, is a reprint of the 3 publications as they appear in medrxiv:

Pagadala, Meghana; Lynch, Julie; Karunamuni, Roshan; Alba, Patrick; Lee, Kyung; Agiri, Fatai; Anglin, Tori; Carter, Hannah; Gaziano, J. Michael; Jasuja, Guneet; Deka, Rishi; Rose, Brent; Panizzon, Matthew; Hauger, Richard; Seibert, Tyler. Evaluating a polygenic hazard score

to predict risk of developing metastatic or fatal prostate cancer in the multi-ancestry Million Veteran Program cohort. The dissertation author was the primary author of this material.

Pagadala, Meghana; Linscott, Joshua; Talwar, James; Seibert, Tyler; Rose, Brent; Lynch, Julie; Panizzon, Matt; Hauger, Richard; Hansen, Moritz; Sammon, Jesse; Hayn, Matthew; Kader, Karim; Carter, Hannah; Ryan, Stephen. PRState: Incorporating Genetic Ancestry in Prostate Cancer Risk scores for African American Men. The dissertation author was the primary author of this material.

Pagadala, Meghana; Jasuja, Guneet; Palnati, Madhuri; Lynch, Julie; Anglin, Tori; Chang, Nelson; Deka, Rishi; Lee, Kyung; Agiri, Fati; Seibert, Tyler; Rose, Brent; Carter, Hannah; Panizzon, Matthew; Hauger, Richard. Discovery of Novel Trans-Ancestry and Ancestry-Specific Gene Loci for Total Testosterone in a Multi-Ancestral Analysis of Men in the Million Veteran Program.

VITA

2012-2016 California Institute of Technology
Research Assistant

2016 California Institute of Technology
Bachelor of Science, Chemistry

2016-2018 University of California San Diego
MD Student

2022 University of California San Diego
Doctor of Philosophy, Biomedical Sciences

PUBLICATIONS

P. Zhang, I. Kitchen-Smith, L. Xiong, G. Stracquadanio, K. Brown, P. H. Richter, M. D. Wallace, E. Bond, N. Sahgal, S. Moore, S. Nornes, S. De Val, M. Surakhy, D. Sims, X. Wang, D. A. Bell, J. Zeron-Medina, Y. Jiang, A. J. Ryan, J. L. Selfe, J. Shipley, S. Kar, P. D. Pharoah, C. Loveday, R. Jansen, L. F. Grochola, C. Palles, A. Protheroe, V. Millar, D. V. Ebner, **M. Pagadala**, S. P. Blagden, T. S. Maughan, E. Domingo, I. Tomlinson, C. Turnbull, H. Carter, G. L. Bond, Germline and Somatic Genetic Variants in the p53 Pathway Interact to Affect Cancer Risk, Progression, and Drug Response. *Cancer Res.* **81**, 1667–1680 (2021).

M. Surakhy, M. Wallace, E. Bond, L. F. Grochola, H. Perez, M. Di Giovannantonio, P. Zhang, D. Malkin, H. Carter, I. Z. S. Parise, G. Zambetti, H. Komechen, M. M. Paraizo, **M. S. Pagadala**, E. M. Pinto, E. Lalli, B. C. Figueiredo, G. L. Bond, A common polymorphism in the retinoic acid pathway modifies adrenocortical carcinoma age-dependent incidence. *Br. J. Cancer*. **122**, 1231–1241 (2020).

Z. Wang, X. Fan, Y. Shen, **M. S. Pagadala**, R. Signer, K. J. Cygan, W. G. Fairbrother, H. Carter, W. K. Chung, K.-L. Huang, Non-cancer-related pathogenic germline variants and expression consequences in ten-thousand cancer genomes. *Genome Med.* **13**, 147 (2021).

D. Nachmanson, **M. Pagadala**, J. Steward, C. Cheung, L. K. Bruce, N. Q. Lee, T. J. O'Keefe, G. Y. Lin, F. Hasteh, G. P. Morris, H. Carter, O. Harismendy, Accurate genome-wide germline profiling from decade-old archival tissue DNA reveals the contribution of common variants to precancer disease outcome, , doi:10.1101/2022.03.31.22273116.

M. S. Pagadala*, J. A. Linscott*, J. Talwar, T. Seibert, B. Rose, J. Lynch, M. Panizzon, R. Hauger, M. H. Hansen, J. D. Sammon, M. H. Hayn, K. Kader, H. Carter, S. T. Ryan, PRState: Incorporating Genetic Ancestry in Prostate Cancer Risk scores for African American Men, , doi:10.1101/2022.03.12.22271020.

M. S. Pagadala, J. Lynch, R. Karunamuni, P. R. Alba, K. M. Lee, F. Y. Agiri, T. Anglin, H. Carter, J. Michael Gaziano, G. K. Jasuja, R. Deka, B. S. Rose, M. S. Panizzon, R. L. Hauger, T.

M. Seibert, Evaluating a polygenic hazard score to predict risk of developing metastatic or fatal prostate cancer in the multi-ancestry Million Veteran Program cohort, , doi:10.1101/2021.09.24.21264093.

J. Talwar, D. Laub, **M. Pagadala**, A. Castro, M. Lewis, G. E. Luebeck, G. P. Morris, R. M. Salem, W. K. Thompson, K. Curtius, M. Zanetti, H. Carter, Autoimmune Alleles at the Major Histocompatibility Locus Modify Melanoma Susceptibility, , doi:10.1101/2021.08.12.456166.

M. Pagadala, V. H. Wu, E. Pérez-Guijarro, H. Kim, A. Castro, J. Talwar, C. Gonzalez-Colin, S. Cao, B. J. Schmiedel, R. M. Salem, G. P. Morris, O. Harismendy, S. P. Patel, J. P. Mesirov, M. Zanetti, C.-P. Day, C. C. Fan, W. K. Thompson, G. Merlino, J. Silvio Gutkind, P. Vijayanand, H. Carter, Germline variants that influence the tumor immune microenvironment also drive response to immunotherapy, , doi:10.1101/2021.04.14.436660.

O. S. Chapman, J. Luebeck, S. Wani, A. Tiwari, **M. Pagadala**, S. Wang, J. D. Larson, J. T. Lange, I. T.-L. Wong, S. R. Dehkordi, S. Chandran, M. Adam, Y. Lin, E. Juarez, J. T. Robinson, S. Sridhar, D. M. Malicki, N. Coufal, M. Levy, J. R. Crawford, S. L. Pomeroy, J. Rich, R. H. Scheuermann, H. Carter, J. Dixon, P. S. Mischel, E. Fraenkel, R. J. Wechsler-Reya, V. Bafna, J. P. Mesirov, L. Chavez, The landscape of extrachromosomal circular DNA in medulloblastoma. *bioRxiv* (2021), p. 2021.10.18.464907.

M. S. Pagadala, G. K. Jasuja, M. Palnati, J. Lynch, T. Anglin, N. Chang, R. Deka, K. M. Lee, F. Y. Agiri, T. M. Seibert, B. S. Rose, H. Carter, M. S. Panizzon*, R. L. Hauger*, Discovery of Novel Trans-Ancestry and Ancestry-Specific Gene Loci for Total Testosterone in a Multi-Ancestral Analysis of Men in the Million Veteran Program, , doi:10.1101/2022.02.16.21265846.

*These authors contributed equally to this work.

ABSTRACT OF THE DISSERTATION

Using Germline Variation to Study Inter-Individual Variability in Cancer Risk and Host Anti-Tumor Immune Response

by

Meghana S. Pagadala

Doctor of Philosophy in Biomedical Sciences

University of California San Diego, 2022

Professor Hannah Carter, Chair
Professor Jill Mesirov, Co-Chair

Cancer is a complex disease driven by genetic variation(*1–3*). Two main types of genetic variation include germline variation, which is inherited, and somatic variation, which is acquired through environmental exposures and endogenous processes, such as DNA replication. Both types of genetic variation have been critical for precision medicine, tailored to each patient's individual cancer. Traditionally, germline variation has been used for risk stratification while somatic variation can be used for treatment selection. For patient risk stratification, several genome-wide

association studies (GWASs) have been conducted to identify underlying genetic determinants that can predict an individual's cancer risk. In the GWAS catalog to date, 128,550 associations and over 4,000 publications have been reported(*4*). Despite the wealth of information provided by GWAS studies, biological insights from these studies are limited and there is still a poor understanding of how inherited variation contributes to cancer behavior.

One major limitation of GWAS studies is the lack of clinical information to understand association. Most large databases only contain basic information, such as demographics and cancer status. However, germline variation can underlie important patterns of tumor behaviors, such as the immune microenvironment, cancer driver mutation frequency and response to therapy. A better understanding of these germline-somatic interactions can help to improve precision medicine efforts. Another major limitation of GWAS studies is lack of diverse genetic cohorts. Genetic cohorts are dominated by individuals of European ancestry and few genetic cohorts of underrepresented populations, such as African-American and Hispanic individuals are exist. In this dissertation, I address these limitations through characterization of germline determinants underlying the tumor immune microenvironment and diverse ancestral populations.

First, in Chapter 1, I identified and characterized germline variants underlying the tumor immune microenvironment in one of the largest adult cancer cohorts, the Cancer Genome Atlas (TCGA). This analysis was motivated by the fact that few biomarkers for immunotherapy exist and a better understanding of germline determinants underlying tumor-immune interactions are needed, especially considering germline variants are present in immune cells along with cancer cells. I identified tumor immune microenvironment SNPs (TIME-SNPs) underlying 157 SNP-heritable immune phenotype components (IP-components) through our own TCGA analysis. Combining these TIME-SNPs with ones collected from literature, we then evaluated cell-type

effects of these variants. Finally, we used various bioinformatic pipelines and databases to determine which TIME-SNPs were implicated in cancer risk, survival and ICB response. We validated one of the genes implicated by our TIME-SNPs as a potential novel immunotherapeutic target.

Next, in Chapter 2, I explored TIME-SNPs in a pediatric cancer cohort collected from multiple databases. Pediatric patients have fewer environmental exposures compared to adults and have traditionally not been good candidates for immunotherapy. Building on analysis from Chapter 1, I processed genotypes from multiple pediatric cancer cohorts and specifically explored TIME-SNPs related to antigen presentation and macrophage infiltration as these were prominent IP components explored in Chapter 1. Also, pediatric cancer patients with mismatch repair (MMR) deficiencies could be potential good candidates for immunotherapy. Thus, we explored germline determinants underlying MMR and their associations with the TIME.

In Chapter 3, I analyzed one of the largest and most diverse genetic databases, the Million Veteran Program, for ancestry-specific associations in testosterone and prostate cancer. Specifically, analyses conducted were 1) a multi-ancestral analysis of total testosterone levels, 2) discovery and evaluation of an African-ancestry specific polygenic risk score and 3) evaluation of a polygenic hazard score for prostate cancer in a multi-ancestry cohort. Through these analyses, I found that genetic associations can differ based on ancestral background and identified novel ancestry-specific associations that improved prostate cancer prediction.

With the culmination of these chapters, I demonstrate that inherited variation underlying the tumor immune microenvironment and diverse populations can improve our understanding of cancer and improve precision medicine efforts.

INTRODUCTION

"Cancer is an expansionist disease; it invades through tissues, sets up colonies in hostile landscapes, seeking 'sanctuary' in one organ and then immigrating to another. It lives desperately, inventively, fiercely, territorially, cannily, and defensively—at times, as if teaching us how to survive. To confront cancer is to encounter a parallel species, one perhaps more adapted to survival than even we are."

–Siddhartha Mukherkee, Emperor of All Maladies

Heritability and Cancer

Many mechanisms underlie conversion of a normal cell to a tumor cell, making cancer an extremely complex disease to treat. At the heart of cancer is genetic variation. Numerous cancer genomics studies have demonstrated that DNA sequence changes cause cancer(*5*, *6*). To cure cancer, a project to sequence the whole human genome was started in 1990 and completed in 2022(*7–10*). Through sequencing the human genome, our understanding of the genetic contribution to cancer deepened. Unfortunately, the global burden of cancer is still enormous, with more than 12 million new cases and 8 million cancer deaths (*11*).

One of the major clinically relevant genetic findings was discovery of heritable cancers. These heritable cancers were seen more frequently in families and thus are often referred to as familial cancers. In familiar cancers, germline mutations in cancer predisposition genes (CPGs) are passed down, significantly increasing risk of cancer in offspring. Prominent cancer predispositions genes, include *TP53, APC, BRCA2, NF1, RB1,* and *RUNX1,* which are related to DNA damage repair. For example, *BRCA1* and *BRCA2* are tumor suppressor genes the produce proteins involved in DNA repair. 55-72% of women with a pathogenic *BRCA1* variant develop breat cancer compared to normal risk frequency of 13% (*12–14*). Fortunately, inherited cancers

are rare and comprise only ~5-10% of cancers. As a result, inherited variation or germline variation was once throught to contribute relatively little to the development of cancer; however, there is evidence that inherited variations contributes to the 90-95% of cancers considered "sporadic cancers." Indeed, twin studies estimate overall heritability of cancer at around 33%, with some tumors showing evidence for >50% heritability(*15*). There is a need to better understand the contribution of inherited variation to all cancers, not just familial cancers (*16*). Furthermore, familial cancers are often driven by rare variants with minor allele frequency (MAF) < 1%; however, understanding the role of common germline variation (MAF > 1%) in cancer is important.

Technologies for Probing the Inherited Genome

SNP arrays are most used for capturing germline variant information due to their cost effectiveness and ease of use. SNP arrays usually are comprised of probes that interrogate common germline SNPs; however, SNP arrays have improved to have a higher density of rare variants and specific disease-relevant SNPs. Custom SNP arrays can be designed; however, most commercial SNP arrays select probes that are common germline markers of a haplotype, or block of SNPs that are inherited together. For example, the OncoArray SNP array is comprised of over 600,000 markers of which half are common germline variant markers while the rest were selected from known cancer GWAS loci(*17*). These probes can then be used to estimate missing SNPs through imputation.

Imputation is process to estimate missing genotypes. The process relies on a reference panel and an input of high-quality genotypes of a population of interest. Most common tools for imputation include Beagle(*18*), IMPUTE2(*19*), MACH+minimac3(*20*),

SHAPEIT2+IMPUTE2(*21*). Imputation uses a hidden markov model (HMM)) to infer missing genotypes based on reference panels, which usually are based on 1000 Genomes project(*22*) or Haplotype Reference Consortium(*23*) if imputing in a European population. To better impute genotypes in non-European individuals, CAAPA panel for African Americans(*24*), Genome Asia for Asian individuals(*25*) and multi-ethnic HLA panel for HLA imputation(*26*) have been published. Most recently, the TOPMed panel built from 97,256 diverse human genomes and 308,107,085 variants outperforms all current reference panels(*27*). Genotyping density, size and type of reference panel and tool used for imputation can all affect results(*28*). After imputation, quality control metrics that take into account imputation accuracy, allele frequency and deviation from Hardy-Weinberg equilibrium are recommended.

Table 0.1 Genetic Cohorts Used for Analysis Genetic cohorts used for thesis chapters 1-3.

Cohort	Description	# individuals
UK Biobank	Longitudinal study of healthy individuals in Europe	500,000
TCGA	Adult cancer patients from 33 different cancer types	10,748
ELLIPSE Consortium	Prostate cancer individuals	91,644
Million Veteran Program	Genetic study of VA participants	
DRIVE Consortium	Breast cancer individuals	60,015
GTEx Consortium	Cell-type specific genetic profiling	15,201
DICE	Immune cell specific eQTL profiling	91

Whole-genome sequencing (WGS) is another more expensive method to acquire germline information. WGS is advantageous in that it is not biased by array probes and thus can fuel novel germline associations. Also, imputation often performs poorly with very rare variants, while WGS can detect these rare variants with enough coverage. In the future, as WGS become more cost-effective, germline variation will most likely be acquired through WGS(*29*). Several consortium and countries, such as Denmark, United Kingdom, United States, are collecting genetic cohorts to serve as resources for germline studies. The cohorts used for this dissertation are detailed below and currently offer SNP array-acquired genotypes. However, many are now sequencing these samples also.

Genome-Wide Association Studies

A genome-wide association study (GWAS) tests associations between millions of genetic variants and a specified phenotype, such as risk of cancer or height. The first GWAS was performed for age-related macular degeneration(*30*) and since then thousand of GWASs have been published. A significance threshold of 5x10-8 is frequently used for statistical cutoffs; however, recent literature suggests that meaningful, biological associations sometimes do not reach this threshold. Thus, a recommended suggestive threshold of 1x10-5 is used in studies also.

More than 50 cancer GWAS studies covering 15 different malinancies have been published, revealing some valuable biological associations. For example, *JAK2* risk variants have been implicated in myeloproliferative neoplasms while *KITLG* risk variants have been linked to testicular cancer. In breast cancer, variants implicating *FGFR2, TOX3, LSP1,* and *STXBP4* have been identified, although their effects have been modest(*31–37*). In prostate cancer, GWAS loci implicate *BRCA2, MSMB, KL3K3* and *KLK2(38–50)*. However, identified loci to date only explain

~20% of familial risk and limited studies explore association with aggressiveness of disease(*51*). Some prominent risk loci, for example, the 8q24 risk loci, do not implicate any genes and their biological link remains a mystery. Methods linking expression with GWAS statistics, such as transcriptome-wide association studies (TWASs), have been helpful in deriving biological meaning; however, more methods to interpret GWASs are needed.

Tumor Immune Microenvironment and Immunotherapy

Part of the complexity of cancer is that it is truly an ecosystem composed of tumor cells, immune cells, stromal cells and extracellular matrix. Immune cells are critical to dictating tumor development and behavior. In an ideal scenario, an individual's immune system can detect and eliminate tumor cells before detection. However, all cancers become good a "hiding" from the immune system, or immune evasion. In a process known as "immunoediting," tumor cells undergo changes in immunogenicity due to host tumor responses. This process consists of: 1) elimination by the immune system, 2) equilibrium where some tumor cell escape the elimination phase and 3) escape where tumor cells grow and expand, successfully evading the immune system. One of the main mechanisms of immune evasion is through downregulation of the antigen presentation machinery, such as major histocompatibility complex (MHC) I, proteosome subunits, transporter associated with antigen processing (TAP) proteins and tapasin(*52–56*). Tumor cells and other TIME cells can also release immunosuppressive molecules such as TNF-alpha, IL-1, IL-6, CSF-1, IL-8, IL-10, VEGF, and TGF-beta, which assist in immune evasion(*57–59*).

Immunotherapy is a promising avenue for arming the immune system to go after tumor cells. Immune checkpoint blockade, which inhibits negative regulators of the immune systems, such as *PD-1, PD-L1* and *CTLA4,* has seen impressive outcomes in melanoma with some results

also observed in non-small cell lung cancer, small-cell lung cancer, renal cell carcinoma and prostate cancer(*60–62*). Despite some promising results with immune checkpoint blockade, generally response rates are low. Biomarkers for immune checkpoint blockade usually are tumor mutational burden (TMB) and immune checkpoint expression. Additionally, other markers such as *CCR5* and *CXCL13* have been identified. However, better biomarkers for response are still needed. Additionally, other forms of immunotherapy such as adoptive cell transfer (ACT) and immune vaccines are under development and may offer alternative options for eliminating cancer.

Genetic background (the germline genome) is responsible for significant interindividual variability of host immune responses(*63*, *64*). Large-scale twin studies found more than 75% of immune phenotypes had major heritable components(*65*). Indeed, GWAS have identified variants associated with leukocyte counts(*66*, *67*), severity of tissue rejection in transplantation and autoimmune diseases(*68*). GWAS of the tumor immune micoenvironment have also been conducted(*69*); however, these studies are complicated by the poor understanding of the tumor immune microenvironment, limited interpretation of non-coding variants and diversity of cells involved. Cell-type specific eQTL discovery and non-coding variant interpretation methods methods have been helpful in making progress towards understanding complex mechanisms of germline variants (*70*, *71*).

Polygenic Risk Scores

While many germline variants are expected to have small effect sizes, genetic risk scores that aggregate the contribution of multiple small effect variants have been effective for identifying individuals at higher risk of certain cancers. Polygenic risk scores (PRSs) are weighted sums of number of risk alleles and an assigned weight, usually an odds ratio for binary traits or beta values

for continuous traits from GWAS statistics. PRS is a useful tool for assessing an individuals genetic liability and have been used in Alzheimer's disease, type 2 diabetes, breast and prostate cancer and cardiovascular disease. A PRS of 303 risk variants was able to achieve an AUC of 0.63 and odds ratio of 0.63 for prediction of development for breast cancer(*72*). Of course, PRS needs to be combined with other information to accurately understand a patient's tumor; however, PRSs can be powerful screening tools for identifying patients at higher risk of cancer(*73*).

Diversity in Germline Studies

Most GWASs are conducted in individuals primarily of European ancestry. Indeed, in the GWAS Catalog, < 3% of study participants were of African ancestry(*74*). The lack of diversity in genetic studies can cause issues as germline variants, linkage disequilibrium patterns and phenotype prevalance can be vastly different in different populations. One major challenge for understanding germline variation in non-European populations is that most initial detection of SNPs are focused on population of European ancestry and many tag SNPs on genotyping panels may not survey the appropriate SNPs needed for a non-European population. This can result in an ascertainment bias, where variants have a comparatively higher minor allele frequency and thus higher expected heterozygosity that other SNPs. Furthermore, admixed individuals present another challenge and might have local ancestral patterns that mediate their disease risk. Some methods are in development to address these issues. Diversity in germline studies is critical as they form the basis for clinical tools that can modify patient care.

References

1. Hallmarks of Cancer: The Next Generation. *Cell*. **144**, 646–674 (2011).

2. The Hallmarks of Cancer. *Cell*. **100**, 57–70 (2000).

3. D. Hanahan, Hallmarks of Cancer: New Dimensions. *Cancer Discovery*. **12** (2022), pp. 31–46.

4. J. MacArthur, E. Bowler, M. Cerezo, L. Gil, P. Hall, E. Hastings, H. Junkins, A. McMahon, A. Milano, J. Morales, Z. M. Pendlington, D. Welter, T. Burdett, L. Hindorff, P. Flicek, F. Cunningham, H. Parkinson, The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45**, D896–D901 (2017).

5. D. Stehelin, H. E. Varmus, J. M. Bishop, P. K. Vogt, DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA. *Nature*. **260** (1976), pp. 170–173.

6. C. J. Tabin, S. M. Bradley, C. I. Bargmann, R. A. Weinberg, A. G. Papageorge, E. M. Scolnick, R. Dhar, D. R. Lowy, E. H. Chang, Mechanism of activation of a human oncogene. *Nature*. **300**, 143–149 (1982).

7. International Human Genome Sequencing Consortium, Finishing the euchromatic sequence of the human genome. *Nature*. **431**, 931–945 (2004).

8. E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, Y. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C.

Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blöcker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowki, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, A. Patrinos, M. J. Morgan, P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, Y. J. Chen, J. Szustakowki, International Human Genome Sequencing Consortium, Initial sequencing and analysis of the human genome. *Nature*. **409**, 860–921 (2001).

9. J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R. R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y. H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigó, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y. H. Chiang, M. Coyne, C. Dahlke, A. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings,

C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, X. Zhu, The sequence of the human genome. *Science*. **291**, 1304–1351 (2001).

10. S. Nurk, S. Koren, A. Rhie, M. Rautiainen, A. V. Bzikadze, A. Mikheenko, M. R. Vollger, N. Altemose, L. Uralsky, A. Gershman, S. Aganezov, S. J. Hoyt, M. Diekhans, G. A. Logsdon, M. Alonge, S. E. Antonarakis, M. Borchers, G. G. Bouffard, S. Y. Brooks, G. V. Caldas, N.-C. Chen, H. Cheng, C.-S. Chin, W. Chow, L. G. de Lima, P. C. Dishuck, R. Durbin, T. Dvorkina, I. T. Fiddes, G. Formenti, R. S. Fulton, A. Fungtammasan, E. Garrison, P. G. S. Grady, T. A. Graves-Lindsay, I. M. Hall, N. F. Hansen, G. A. Hartley, M. Haukness, K. Howe, M. W. Hunkapiller, C. Jain, M. Jain, E. D. Jarvis, P. Kerpedjiev, M. Kirsche, M. Kolmogorov, J. Korlach, M. Kremitzki, H. Li, V. V. Maduro, T. Marschall, A. M. McCartney, J. McDaniel, D. E. Miller, J. C. Mullikin, E. W. Myers, N. D. Olson, B. Paten, P. Peluso, P. A. Pevzner, D. Porubsky, T. Potapova, E. I. Rogaev, J. A. Rosenfeld, S. L. Salzberg, V. A. Schneider, F. J. Sedlazeck, K. Shafin, C. J. Shew, A. Shumate, Y. Sims, A. F. A. Smit, D. C. Soto, I. Sović, J. M. Storer, A. Streets, B. A. Sullivan, F. Thibaud-Nissen, J. Torrance, J. Wagner, B. P. Walenz, A. Wenger, J. M. D. Wood, C. Xiao, S. M. Yan, A. C. Young, S. Zarate, U. Surti, R. C. McCoy, M. Y. Dennis, I. A. Alexandrov, J. L. Gerton, R. J. O'Neill, W. Timp, J. M. Zook, M. C. Schatz, E. E. Eichler, K. H. Miga, A. M. Phillippy, The complete sequence of a human genome. *Science*. **376**, 44–53 (2022).

11. J. Ferlay, I. Soerjomataram, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. M. Parkin, D. Forman, F. Bray, Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int. J. Cancer*. **136**, E359–86 (2015).

12. K. B. Kuchenbaecker, J. L. Hopper, D. R. Barnes, K.-A. Phillips, T. M. Mooij, M.-J. Roos-Blom, S. Jervis, F. E. van Leeuwen, R. L. Milne, N. Andrieu, D. E. Goldgar, M. B. Terry, M. A. Rookus, D. F. Easton, A. C. Antoniou, BRCA1 and BRCA2 Cohort Consortium, L. McGuffog, D. G. Evans, D. Barrowdale, D. Frost, J. Adlard, K.-R. Ong, L. Izatt, M. Tischkowitz, R. Eeles, R. Davidson, S. Hodgson, S. Ellis, C. Nogues, C. Lasset, D. Stoppa-Lyonnet, J.-P. Fricker, L. Faivre, P. Berthet, M. J. Hooning, L. E. van der Kolk, C. M. Kets, M. A. Adank, E. M. John, W. K. Chung, I. L. Andrulis, M. Southey, M. B. Daly, S. S. Buys, A. Osorio, C. Engel, K. Kast, R. K. Schmutzler, T. Caldes, A. Jakubowska, J. Simard, M. L. Friedlander, S.-A. McLachlan, E. Machackova, L. Foretova, Y. Y. Tan, C. F. Singer, E. Olah, A.-M. Gerdes, B. Arver, H. Olsson, Risks of Breast, Ovarian, and Contralateral Breast Cancer for BRCA1 and BRCA2 Mutation Carriers. *JAMA*. **317**, 2402–2416 (2017).

13. A. Antoniou, P. D. P. Pharoah, S. Narod, H. A. Risch, J. E. Eyfjord, J. L. Hopper, N. Loman, H. Olsson, O. Johannsson, Å. Borg, B. Pasini, P. Radice, S. Manoukian, D. M. Eccles, N. Tang, E. Olah, H. Anton-Culver, E. Warner, J. Lubinski, J. Gronwald, B. Gorski, H. Tulinius, S. Thorlacius, H. Eerola, H. Nevanlinna, K. Syrjäkoski, O.-P. Kallioniemi, D. Thompson, C. Evans, J. Peto, F. Lalloo, D. G. Evans, D. F. Easton, Average Risks of Breast and Ovarian Cancer Associated with BRCA1 or BRCA2 Mutations Detected in Case Series Unselected for Family History: A Combined Analysis of 22 Studies. *The American Journal of Human Genetics*. **72** (2003), pp. 1117–1130.

14. S. Chen, G. Parmigiani, Meta-analysis of BRCA1 and BRCA2 penetrance. *J. Clin. Oncol.* **25**, 1329–1333 (2007).

15. L. A. Mucci, J. B. Hjelmborg, J. R. Harris, K. Czene, D. J. Havelick, T. Scheike, R. E. Graff, K. Holst, S. Möller, R. H. Unger, C. McIntosh, E. Nuttall, I. Brandt, K. L. Penney, M. Hartman, P. Kraft, G. Parmigiani, K. Christensen, M. Koskenvuo, N. V. Holm, K. Heikkilä, E. Pukkala, A. Skytthe, H.-O. Adami, J. Kaprio, Nordic Twin Study of Cancer (NorTwinCan) Collaboration, Familial risk and heritability of cancer among twins in Nordic countries. *JAMA*. **315**, 68–76 (2016).

16. E. Cano-Gamez, G. Trynka, From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. *Front. Genet.* **11**, 424 (2020).

17. C. I. Amos, J. Dennis, Z. Wang, J. Byun, F. R. Schumacher, S. A. Gayther, G. Casey, D. J. Hunter, T. A. Sellers, S. B. Gruber, A. M. Dunning, K. Michailidou, L. Fachal, K. Doheny, A. B. Spurdle, Y. Li, X. Xiao, J. Romm, E. Pugh, G. A. Coetzee, D. J. Hazelett, S. E. Bojesen, C. Caga-Anan, C. A. Haiman, A. Kamal, C. Luccarini, D. Tessier, D. Vincent, F. Bacot, D. J. Van Den Berg, S. Nelson, S. Demetriades, D. E. Goldgar, F. J. Couch, J. L. Forman, G. G. Giles, D. V. Conti, H. Bickeböller, A. Risch, M. Waldenberger, I. Brüske-Hohlfeld, B. D. Hicks, H. Ling, L. McGuffog, A. Lee, K. Kuchenbaecker, P. Soucy, J. Manz, J. M. Cunningham, K. Butterbach, Z. Kote-Jarai, P. Kraft, L. FitzGerald, S. Lindström, M. Adams, J. D. McKay, C. M. Phelan, S. Benlloch, L. E. Kelemen, P. Brennan, M. Riggan, T. A. O'Mara, H. Shen, Y. Shi, D. J. Thompson, M. T. Goodman, S. F. Nielsen, A. Berchuck, S. Laboissiere, S. L. Schmit, T. Shelford, C. K. Edlund, J. A. Taylor, J. K. Field, S. K. Park, K. Offit, M. Thomassen, R. Schmutzler, L. Ottini, R. J. Hung, J. Marchini, A. Amin Al Olama, U. Peters, R. A. Eeles, M. F. Seldin, E. Gillanders, D. Seminara, A. C. Antoniou, P. D. P. Pharoah, G. Chenevix-Trench, S. J. Chanock, J. Simard, D. F. Easton, The OncoArray Consortium: A Network for Understanding the Genetic Architecture of Common Cancers. *Cancer Epidemiol. Biomarkers Prev.* **26**, 126–135 (2017).

18. B. L. Browning, X. Tian, Y. Zhou, S. R. Browning, Fast two-stage phasing of large-scale sequence data. *Am. J. Hum. Genet.* **108**, 1880–1890 (2021).

19. B. N. Howie, P. Donnelly, J. Marchini, A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5**, e1000529 (2009).

20. S. Das, L. Forer, S. Schönherr, C. Sidore, A. E. Locke, A. Kwong, S. I. Vrieze, E. Y. Chew, S. Levy, M. McGue, D. Schlessinger, D. Stambolian, P.-R. Loh, W. G. Iacono, A. Swaroop, L. J. Scott, F. Cucca, F. Kronenberg, M. Boehnke, G. R. Abecasis, C. Fuchsberger, Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).

21. O. Delaneau, J. Marchini, J.-F. Zagury, A linear complexity phasing method for thousands of genomes. *Nat. Methods*. **9**, 179–181 (2011).

22. The 1000 Genomes Project Consortium, A. Auton, G. R. Abecasis, D. M. Altshuler (Co-Chair), R. M. Durbin (Co-Chair), G. R. Abecasis, D. R. Bentley, A. Chakravarti, A. G. Clark, P. Donnelly, E. E. Eichler, P. Flicek, S. B. Gabriel, R. A. Gibbs, E. D. Green, M. E. Hurles, B. M. Knoppers, J. O. Korbel, E. S. Lander, C. Lee, H. Lehrach, E. R. Mardis, G. T. Marth,

G. A. McVean, D. A. Nickerson, J. P. Schmidt, S. T. Sherry, J. Wang, R. K. Wilson, R. A. Gibbs (Principal, E. Boerwinkle, H. Doddapaneni, Y. Han, V. Korchina, C. Kovar, S. Lee, D. Muzny, J. G. Reid, Y. Zhu, J. Wang (Principal, Y. Chang, Q. Feng, X. Fang, X. Guo, M. Jian, H. Jiang, X. Jin, T. Lan, G. Li, J. Li, Y. Li, S. Liu, X. Liu, Y. Lu, X. Ma, M. Tang, B. Wang, G. Wang, H. Wu, R. Wu, X. Xu, Y. Yin, D. Zhang, W. Zhang, J. Zhao, M. Zhao, X. Zheng, E. S. Lander (Principal, D. M. Altshuler, S. B. Gabriel (Co-Chair), N. Gupta, N. Gharani, L. H. Toji, N. P. Gerry, A. M. Resch, P. Flicek (Principal, J. Barker, L. Clarke, L. Gil, S. E. Hunt, G. Kelman, E. Kulesha, R. Leinonen, W. M. McLaren, R. Radhakrishnan, A. Roa, D. Smirnov, R. E. Smith, I. Streeter, A. Thormann, I. Toneva, B. Vaughan, X. Zheng-Bradley, D. R. Bentley (Principal, R. Grocock, S. Humphray, T. James, Z. Kingsbury, H. Lehrach (Principal, R. Sudbrak (Project L, M. W. Albrecht, V. S. Amstislavskiy, T. A. Borodina, M. Lienhard, F. Mertes, M. Sultan, B. Timmermann, M.-L. Yaspo, E. R. Mardis (Co-Princi, R. K. Wilson (Co-Princi, L. Fulton, R. Fulton, S. T. Sherry (Principal, V. Ananiev, Z. Belaia, D. Beloslyudtsev, N. Bouk, C. Chen, D. Church, R. Cohen, C. Cook, J. Garner, T. Hefferon, M. Kimelman, C. Liu, J. Lopez, P. Meric, C. O'Sullivan, Y. Ostapchuk, L. Phan, S. Ponomarov, V. Schneider, E. Shekhtman, K. Sirotkin, D. Slotta, H. Zhang, G. A. McVean (Principal, R. M. Durbin (Principal, S. Balasubramaniam, J. Burton, P. Danecek, T. M. Keane, A. Kolb-Kokocinski, S. McCarthy, J. Stalker, M. Quail, J. P. Schmidt (Principal, C. J. Davies, J. Gollub, T. Webster, B. Wong, Y. Zhan, A. Auton (Principal, C. L. Campbell, Y. Kong, A. Marcketta, R. A. Gibbs (Principal, F. Yu (Project L, L. Antunes, M. Bainbridge, D. Muzny, A. Sabo, Z. Huang, J. Wang (Principal, L. J. M. Coin, L. Fang, X. Guo, X. Jin, G. Li, Q. Li, Y. Li, Z. Li, H. Lin, B. Liu, R. Luo, H. Shao, Y. Xie, C. Ye, C. Yu, F. Zhang, H. Zheng, H. Zhu, C. Alkan, E. Dal, F. Kahveci, G. T. Marth (Principal, E. P. Garrison (Project L, D. Kural, W.-P. Lee, W. Fung Leong, M. Stromberg, A. N. Ward, J. Wu, M. Zhang, M. J. Daly (Principal, M. A. DePristo (Project L, R. E. Handsaker (Project L, D. M. Altshuler, E. Banks, G. Bhatia, G. del Angel, S. B. Gabriel, G. Genovese, N. Gupta, H. Li, S. Kashin, E. S. Lander, S. A. McCarroll, J. C. Nemesh, R. E. Poplin, S. C. Yoon (Principal, J. Lihm, V. Makarov, A. G. Clark (Principal, S. Gottipati, A. Keinan, J. L. Rodriguez-Flores, J. O. Korbel (Principal, T. Rausch (Project L, M. H. Fritz, A. M. Stütz, P. Flicek (Principal, K. Beal, L. Clarke, A. Datta, J. Herrero, W. M. McLaren, G. R. S. Ritchie, R. E. Smith, D. Zerbino, X. Zheng-Bradley, P. C. Sabeti (Principal, I. Shlyakhter, S. F. Schaffner, J. Vitti, D. N. Cooper (Principal, E. V. Ball, P. D. Stenson, D. R. Bentley (Principal, B. Barnes, M. Bauer, R. Keira Cheetham, A. Cox, M. Eberle, S. Humphray, S. Kahn, L. Murray, J. Peden, R. Shaw, E. E. Kenny (Principal, M. A. Batzer (Principal, M. K. Konkel, J. A. Walker, D. G. MacArthur (Principal, M. Lek, R. Sudbrak (Project L, V. S. Amstislavskiy, R. Herwig, E. R. Mardis (Co-Princi, L. Ding, D. C. Koboldt, D. Larson, K. Ye, S. Gravel, A. Swaroop, E. Chew, T. Lappalainen (Principal, Y. Erlich (Principal, M. Gymrek, T. Frederick Willems, J. T. Simpson, M. D. Shriver (Principal, J. A. Rosenfeld (Principal, C. D. Bustamante (Principal, S. B. Montgomery (Principal, F. M. De La Vega (Principal, J. K. Byrnes, A. W. Carroll, M. K. DeGorter, P. Lacroute, B. K. Maples, A. R. Martin, A. Moreno-Estrada, S. S. Shringarpure, F. Zakharia, E. Halperin (Principal, Y. Baran, C. Lee (Principal, E. Cerveira, J. Hwang, A. Malhotra (Co-Projec, D. Plewczynski, K. Radew, M. Romanovitch, C. Zhang (Co-Projec, F. C. L. Hyland, D. W. Craig (Principal, A. Christoforides, N. Homer, T. Izatt, A. A. Kurdoglu, S. A. Sinari, K. Squire, S. T. Sherry (Principal, C. Xiao, J. Sebat (Principal, D. Antaki, M. Gujral, A. Noor, K. Ye, E. G. Burchard (Principal, R. D. Hernandez (Principal, C. R. Gignoux, D. Haussler (Principal, S. J. Katzman, W. James Kent, B. Howie, A. Ruiz-Linares

(Principal, E. T. Dermitzakis (Principal, S. E. Devine (Principal, G. R. Abecasis (Principal, H. Min Kang (Project L, J. M. Kidd (Principal, T. Blackwell, S. Caron, W. Chen, S. Emery, L. Fritsche, C. Fuchsberger, G. Jun, B. Li, R. Lyons, C. Scheller, C. Sidore, S. Song, E. Sliwerska, D. Taliun, A. Tan, R. Welch, M. Kate Wing, X. Zhan, P. Awadalla (Principal, A. Hodgkinson, Y. Li, X. Shi (Principal, A. Quitadamo, G. Lunter (Principal, G. A. McVean (Principal, J. L. Marchini (Principal, S. Myers (Principal, C. Churchhouse, O. Delaneau, A. Gupta-Hinch, W. Kretzschmar, Z. Iqbal, I. Mathieson, A. Menelaou, A. Rimmer, D. K. Xifara, T. K. Oleksyk (Principal, Y. Fu (Principal, X. Liu, M. Xiong, L. Jorde (Principal, D. Witherspoon, J. Xing, E. E. Eichler (Principal, B. L. Browning (Principal, S. R. Browning (Principal, F. Hormozdiari, P. H. Sudmant, E. Khurana (Principal, R. M. Durbin (Principal, M. E. Hurles (Principal, C. Tyler-Smith (Principal, C. A. Albers, Q. Ayub, S. Balasubramaniam, Y. Chen, V. Colonna, P. Danecek, L. Jostins, T. M. Keane, S. McCarthy, K. Walter, Y. Xue, M. B. Gerstein (Principal, A. Abyzov, S. Balasubramanian, J. Chen, D. Clarke, Y. Fu, A. O. Harmanci, M. Jin, D. Lee, J. Liu, X. Jasmine Mu, J. Zhang, Y. Zhang, Y. Li, R. Luo, H. Zhu, C. Alkan, E. Dal, F. Kahveci, G. T. Marth (Principal, E. P. Garrison, D. Kural, W.-P. Lee, A. N. Ward, J. Wu, M. Zhang, S. A. McCarroll (Principal, R. E. Handsaker (Project L, D. M. Altshuler, E. Banks, G. del Angel, G. Genovese, C. Hartl, H. Li, S. Kashin, J. C. Nemesh, K. Shakir, S. C. Yoon (Principal, J. Lihm, V. Makarov, J. Degenhardt, J. O. Korbel (Principal, M. H. Fritz, S. Meiers, B. Raeder, T. Rausch, A. M. Stütz, P. Flicek (Principal, F. Paolo Casale, L. Clarke, R. E. Smith, O. Stegle, X. Zheng-Bradley, D. R. Bentley (Principal, B. Barnes, R. Keira Cheetham, M. Eberle, S. Humphray, S. Kahn, L. Murray, R. Shaw, E.-W. Lameijer, M. A. Batzer (Principal, M. K. Konkel, J. A. Walker, L. Ding (Principal, I. Hall, K. Ye, P. Lacroute, C. Lee (Principal, E. Cerveira, A. Malhotra, J. Hwang, D. Plewczynski, K. Radew, M. Romanovitch, C. Zhang, D. W. Craig (Principal, N. Homer, D. Church, C. Xiao, J. Sebat (Principal, D. Antaki, V. Bafna, J. Michaelson, K. Ye, S. E. Devine (Principal, E. J. Gardner (Project L, G. R. Abecasis (Principal, J. M. Kidd (Principal, R. E. Mills (Principal, G. Dayama, S. Emery, G. Jun, X. Shi (Principal, A. Quitadamo, G. Lunter (Principal, G. A. McVean (Principal, K. Chen (Principle, X. Fan, Z. Chong, T. Chen, D. Witherspoon, J. Xing, E. E. Eichler (Principal, M. J. Chaisson, F. Hormozdiari, J. Huddleston, M. Malig, B. J. Nelson, P. H. Sudmant, N. F. Parrish, E. Khurana (Principal, M. E. Hurles (Principal, B. Blackburne, S. J. Lindsay, Z. Ning, K. Walter, Y. Zhang, M. B. Gerstein (Principal, A. Abyzov, J. Chen, D. Clarke, H. Lam, X. Jasmine Mu, C. Sisu, J. Zhang, Y. Zhang, R. A. Gibbs (Principal, F. Yu (Project L, M. Bainbridge, D. Challis, U. S. Evani, C. Kovar, J. Lu, D. Muzny, U. Nagaswamy, J. G. Reid, A. Sabo, J. Yu, X. Guo, W. Li, Y. Li, R. Wu, G. T. Marth (Principal, E. P. Garrison, W. Fung Leong, A. N. Ward, G. del Angel, M. A. DePristo, S. B. Gabriel, N. Gupta, C. Hartl, R. E. Poplin, A. G. Clark (Principal, J. L. Rodriguez-Flores, P. Flicek (Principal, L. Clarke, R. E. Smith, X. Zheng-Bradley, D. G. MacArthur (Principal, E. R. Mardis (Principal, R. Fulton, D. C. Koboldt, S. Gravel, C. D. Bustamante (Principal, D. W. Craig (Principal, A. Christoforides, N. Homer, T. Izatt, S. T. Sherry (Principal, C. Xiao, E. T. Dermitzakis (Principal, G. R. Abecasis (Principal, H. Min Kang, G. A. McVean (Principal, M. B. Gerstein (Principal, S. Balasubramanian, L. Habegger, H. Yu (Principal, P. Flicek (Principal, L. Clarke, F. Cunningham, I. Dunham, D. Zerbino, X. Zheng-Bradley, K. Lage (Principal, J. Berg Jespersen, H. Horn, S. B. Montgomery (Principal, M. K. DeGorter, E. Khurana (Principal, C. Tyler-Smith (Principal, Y. Chen, V. Colonna, Y. Xue, M. B. Gerstein (Principal, S. Balasubramanian, Y. Fu, D. Kim, A. Auton (Principal, A. Marcketta, R. Desalle, A.

Narechania, M. A. Wilson Sayres, E. P. Garrison, R. E. Handsaker, S. Kashin, S. A. McCarroll, J. L. Rodriguez-Flores, P. Flicek (Principal, L. Clarke, X. Zheng-Bradley, Y. Erlich, M. Gymrek, T. Frederick Willems, C. D. Bustamante (Principal, F. L. Mendez, G. David Poznik, P. A. Underhill, C. Lee, E. Cerveira, A. Malhotra, M. Romanovitch, C. Zhang, G. R. Abecasis (Principal, L. Coin (Principal, H. Shao, D. Mittelman, C. Tyler-Smith (Principal, Q. Ayub, R. Banerjee, M. Cerezo, Y. Chen, T. W. Fitzgerald, S. Louzada, A. Massaia, S. McCarthy, G. R. Ritchie, Y. Xue, F. Yang, R. A. Gibbs (Principal, C. Kovar, D. Kalra, W. Hale, D. Muzny, J. G. Reid, J. Wang (Principal, X. Dan, X. Guo, G. Li, Y. Li, C. Ye, X. Zheng, D. M. Altshuler, P. Flicek (Principal, L. Clarke (Project L, X. Zheng-Bradley, D. R. Bentley (Principal, A. Cox, S. Humphray, S. Kahn, R. Sudbrak (Project L, M. W. Albrecht, M. Lienhard, D. Larson, D. W. Craig (Principal, T. Izatt, A. A. Kurdoglu, S. T. Sherry (Principal, C. Xiao, D. Haussler (Principal, G. R. Abecasis (Principal, G. A. McVean (Principal, R. M. Durbin (Principal, S. Balasubramaniam, T. M. Keane, S. McCarthy, J. Stalker, A. Chakravarti (Co-Chair), B. M. Knoppers (Co-Chair), G. R. Abecasis, K. C. Barnes, C. Beiswanger, E. G. Burchard, C. D. Bustamante, H. Cai, H. Cao, R. M. Durbin, N. P. Gerry, N. Gharani, R. A. Gibbs, C. R. Gignoux, S. Gravel, B. Henn, D. Jones, L. Jorde, J. S. Kaye, A. Keinan, A. Kent, A. Kerasidou, Y. Li, R. Mathias, G. A. McVean, A. Moreno-Estrada, P. N. Ossorio, M. Parker, A. M. Resch, C. N. Rotimi, Charmaine D. Royal, K. Sandoval, Y. Su, R. Sudbrak, Z. Tian, S. Tishkoff, L. H. Toji, C. Tyler-Smith, M. Via, Y. Wang, H. Yang, L. Yang, J. Zhu, W. Bodmer, G. Bedoya, A. Ruiz-Linares, Z. Cai, Y. Gao, J. Chu, L. Peltonen, A. Garcia-Montero, A. Orfao, J. Dutil, J. C. Martinez-Cruzado, T. K. Oleksyk, K. C. Barnes, R. A. Mathias, A. Hennis, H. Watson, C. McKenzie, F. Qadri, R. LaRocque, P. C. Sabeti, J. Zhu, X. Deng, P. C. Sabeti, D. Asogun, O. Folarin, C. Happi, O. Omoniwa, M. Stremlau, R. Tariyal, M. Jallow, F. Sisay Joof, T. Corrah, K. Rockett, D. Kwiatkowski, J. Kooner, T. Tịnh Hiê`n, S. J. Dunstan, N. Thuy Hang, R. Fonnie, R. Garry, L. Kanneh, L. Moses, P. C. Sabeti, J. Schieffelin, D. S. Grant, C. Gallo, G. Poletti, D. Saleheen, A. Rasheed, L. D. Brooks, A. L. Felsenfeld, J. E. McEwen, Y. Vaydylevich, E. D. Green, A. Duncanson, M. Dunn, J. A. Schloss, J. Wang, H. Yang, A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison, H. Min Kang, J. O. Korbel, J. L. Marchini, S. McCarthy, G. A. McVean, G. R. Abecasis, Corresponding authors, Steering committee, Production group, Baylor College of Medicine, BGI-Shenzhen, Broad Institute of MIT and Harvard, Coriell Institute for Medical Research, European Molecular Biology Laboratory, European Bioinformatics Institute, Illumina, Max Planck Institute for Molecular Genetics, McDonnell Genome Institute at Washington University, US National Institutes of Health, University of Oxford, Wellcome Trust Sanger Institute, Analysis group, Affymetrix, Albert Einstein College of Medicine, Baylor College of Medicine, BGI-Shenzhen, Bilkent University, Boston College, Broad Institute of MIT and Harvard, Cold Spring Harbor Laboratory, Cornell University, European Molecular Biology Laboratory, European Molecular Biology Laboratory, European Bioinformatics Institute, Harvard University, Human Gene Mutation Database, Illumina, Icahn School of Medicine at Mount Sinai, Louisiana State University, Massachusetts General Hospital, Max Planck Institute for Molecular Genetics, McDonnell Genome Institute at Washington University, McGill University, National Eye Institute, NIH, New York Genome Center, Ontario Institute for Cancer Research, Pennsylvania State University, Rutgers Cancer Institute of New Jersey, Stanford University, Tel-Aviv University, The Jackson Laboratory for Genomic Medicine, T. F. Scientific, Translational Genomics Research Institute, US National Institutes of Health, University of California, San

Diego, University of California, San Francisco, University of California, Santa Cruz, University of Chicago, University College London, University of Geneva, University of Maryland School of Medicine, University of Michigan, University of Montréal, University of North Carolina at Chapel Hill, University of North Carolina at Charlotte, University of Oxford, University of Puerto Rico, University of Texas Health Sciences Center at Houston, University of Utah, University of Washington, Weill Cornell Medical College, Wellcome Trust Sanger Institute, Yale University, Structural variation group, BGI-Shenzhen, Bilkent University, Boston College, Broad Institute of MIT and Harvard, Cold Spring Harbor Laboratory, Cornell University, European Molecular Biology Laboratory, European Molecular Biology Laboratory, European Bioinformatics Institute, Illumina, Leiden University Medical Center, Louisiana State University, McDonnell Genome Institute at Washington University, Stanford University, The Jackson Laboratory for Genomic Medicine, Translational Genomics Research Institute, US National Institutes of Health, University of California, San Diego, University of Maryland School of Medicine, University of Michigan, University of North Carolina at Charlotte, University of Oxford, University of Texas MD Anderson Cancer Center, University of Utah, University of Washington, Vanderbilt University School of Medicine, Weill Cornell Medical College, Wellcome Trust Sanger Institute, Yale University, Exome group, Baylor College of Medicine, BGI-Shenzhen, Boston College, Broad Institute of MIT and Harvard, Cornell University, European Molecular Biology Laboratory, European Bioinformatics Institute, Massachusetts General Hospital, McDonnell Genome Institute at Washington University, McGill University, Stanford University, Translational Genomics Research Institute, US National Institutes of Health, University of Geneva, University of Michigan, University of Oxford, Yale University, Functional interpretation group, Cornell University, European Molecular Biology Laboratory, European Bioinformatics Institute, Harvard University, Stanford University, Weill Cornell Medical College, Wellcome Trust Sanger Institute, Yale University, Chromosome Y group, Albert Einstein College of Medicine, American Museum of Natural History, Arizona State University, Boston College, Broad Institute of MIT and Harvard, Cornell University, European Molecular Biology Laboratory, European Bioinformatics Institute, New York Genome Center, Stanford University, The Jackson Laboratory for Genomic Medicine, University of Michigan, University of Queensland, Virginia Bioinformatics Institute, Wellcome Trust Sanger Institute, Data coordination center group, Baylor College of Medicine, BGI-Shenzhen, Broad Institute of MIT and Harvard, European Molecular Biology Laboratory, European Bioinformatics Institute, Illumina, Max Planck Institute for Molecular Genetics, McDonnell Genome Institute at Washington University, Translational Genomics Research Institute, US National Institutes of Health, University of California, Santa Cruz, University of Michigan, University of Oxford, Wellcome Trust Sanger Institute, Samples and ELSI group, Sample collection, British from England and Scotland (GBR), Colombians in Medellín, Colombia (CLM), Han Chinese South (CHS), Finnish in Finland (FIN), Iberian Populations in Spain (IBS), Puerto Ricans in Puerto Rico (PUR), African Caribbean in Barbados (ACB), Bengali in Bangladesh (BEB), Chinese Dai in Xishuangbanna, China (CDX), Esan in Nigeria (ESN), Gambian in Western Division - Mandinka (GWD), Indian Telugu in the UK (ITU) and Sri Lankan Tamil in the UK (STU), Kinh in Ho Chi Minh City, Vietnam (KHV), Mende in Sierra Leone (MSL), Peruvian in Lima, Peru (PEL), Punjabi in Lahore, Pakistan (PJL), Scientific management, Writing group, A global reference for human genetic variation. *Nature*. **526**, 68–74 (2015).

23. S. McCarthy, S. Das, W. Kretzschmar, O. Delaneau, A. R. Wood, A. Teumer, H. M. Kang, C. Fuchsberger, P. Danecek, K. Sharp, Y. Luo, C. Sidore, A. Kwong, N. Timpson, S. Koskinen, S. Vrieze, L. J. Scott, H. Zhang, A. Mahajan, J. Veldink, U. Peters, C. Pato, C. M. van Duijn, C. E. Gillies, I. Gandin, M. Mezzavilla, A. Gilly, M. Cocca, M. Traglia, A. Angius, J. C. Barrett, D. Boomsma, K. Branham, G. Breen, C. M. Brummett, F. Busonero, H. Campbell, A. Chan, S. Chen, E. Chew, F. S. Collins, L. J. Corbin, G. D. Smith, G. Dedoussis, M. Dorr, A.-E. Farmaki, L. Ferrucci, L. Forer, R. M. Fraser, S. Gabriel, S. Levy, L. Groop, T. Harrison, A. Hattersley, O. L. Holmen, K. Hveem, M. Kretzler, J. C. Lee, M. McGue, T. Meitinger, D. Melzer, J. L. Min, K. L. Mohlke, J. B. Vincent, M. Nauck, D. Nickerson, A. Palotie, M. Pato, N. Pirastu, M. McInnis, J. B. Richards, C. Sala, V. Salomaa, D. Schlessinger, S. Schoenherr, P. E. Slagboom, K. Small, T. Spector, D. Stambolian, M. Tuke, J. Tuomilehto, L. H. Van den Berg, W. Van Rheenen, U. Volker, C. Wijmenga, D. Toniolo, E. Zeggini, P. Gasparini, M. G. Sampson, J. F. Wilson, T. Frayling, P. I. W. de Bakker, M. A. Swertz, S. McCarroll, C. Kooperberg, A. Dekker, D. Altshuler, C. Willer, W. Iacono, S. Ripatti, N. Soranzo, K. Walter, A. Swaroop, F. Cucca, C. A. Anderson, R. M. Myers, M. Boehnke, M. I. McCarthy, R. Durbin, Haplotype Reference Consortium, A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).

24. J. O'Connell, T. Yun, M. Moreno, H. Li, N. Litterman, A. Kolesnikov, E. Noblin, P.-C. Chang, A. Shastri, E. H. Dorfman, S. Shringarpure, 23andMe Research Team, A. Auton, A. Carroll, C. Y. McLean, A population-specific reference panel for improved genotype imputation in African Americans. *Commun Biol*. **4**, 1269 (2021).

25. GenomeAsia100K Consortium, The GenomeAsia 100K Project enables genetic discoveries across Asia. *Nature*. **576**, 106–111 (2019).

26. Y. Luo, M. Kanai, W. Choi, X. Li, K. Yamamoto, K. Ogawa, M. Gutierrez-Arcelus, P. K. Gregersen, P. E. Stuart, J. T. Elder, J. Fellay, M. Carrington, D. W. Haas, X. Guo, N. D. Palmer, Y.-D. I. Chen, J. I. Rotter, K. D. Taylor, S. S. Rich, A. Correa, J. G. Wilson, S. Kathiresan, M. H. Cho, A. Metspalu, T. Esko, Y. Okada, B. Han, P. J. McLaren, S. Raychaudhuri, NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, A high-resolution HLA reference panel capturing global population diversity enables multi-ethnic fine-mapping in HIV host response, , doi:10.1101/2020.07.16.20155606.

27. D. Taliun, D. N. Harris, M. D. Kessler, J. Carlson, Z. A. Szpiech, R. Torres, S. A. G. Taliun, A. Corvelo, S. M. Gogarten, H. M. Kang, A. N. Pitsillides, J. LeFaive, S.-B. Lee, X. Tian, B. L. Browning, S. Das, A.-K. Emde, W. E. Clarke, D. P. Loesch, A. C. Shetty, T. W. Blackwell, A. V. Smith, Q. Wong, X. Liu, M. P. Conomos, D. M. Bobo, F. Aguet, C. Albert, A. Alonso, K. G. Ardlie, D. E. Arking, S. Aslibekyan, P. L. Auer, J. Barnard, R. G. Barr, L. Barwick, L. C. Becker, R. L. Beer, E. J. Benjamin, L. F. Bielak, J. Blangero, M. Boehnke, D. W. Bowden, J. A. Brody, E. G. Burchard, B. E. Cade, J. F. Casella, B. Chalazan, D. I. Chasman, Y.-D. I. Chen, M. H. Cho, S. H. Choi, M. K. Chung, C. B. Clish, A. Correa, J. E. Curran, B. Custer, D. Darbar, M. Daya, M. de Andrade, D. L. DeMeo, S. K. Dutcher, P. T. Ellinor, L. S. Emery, C. Eng, D. Fatkin, T. Fingerlin, L. Forer, M. Fornage, N. Franceschini, C. Fuchsberger, S. M. Fullerton, S. Germer, M. T. Gladwin, D. J. Gottlieb, X. Guo, M. E. Hall, J. He, N. L. Heard-Costa, S. R. Heckbert, M. R. Irvin, J. M. Johnsen, A. D. Johnson, R. Kaplan, S. L. R. Kardia, T. Kelly, S. Kelly, E. E. Kenny, D. P. Kiel, R. Klemmer, B. A. Konkle, C. Kooperberg, A.

Köttgen, L. A. Lange, J. Lasky-Su, D. Levy, X. Lin, K.-H. Lin, C. Liu, R. J. F. Loos, L. Garman, R. Gerszten, S. A. Lubitz, K. L. Lunetta, A. C. Y. Mak, A. Manichaikul, A. K. Manning, R. A. Mathias, D. D. McManus, S. T. McGarvey, J. B. Meigs, D. A. Meyers, J. L. Mikulla, M. A. Minear, B. D. Mitchell, S. Mohanty, M. E. Montasser, C. Montgomery, A. C. Morrison, J. M. Murabito, A. Natale, P. Natarajan, S. C. Nelson, K. E. North, J. R. O'Connell, N. D. Palmer, N. Pankratz, G. M. Peloso, P. A. Peyser, J. Pleiness, W. S. Post, B. M. Psaty, D. C. Rao, S. Redline, A. P. Reiner, D. Roden, J. I. Rotter, I. Ruczinski, C. Sarnowski, S. Schoenherr, D. A. Schwartz, J.-S. Seo, S. Seshadri, V. A. Sheehan, W. H. Sheu, M. B. Shoemaker, N. L. Smith, J. A. Smith, N. Sotoodehnia, A. M. Stilp, W. Tang, K. D. Taylor, M. Telen, T. A. Thornton, R. P. Tracy, D. J. Van Den Berg, R. S. Vasan, K. A. Viaud-Martinez, S. Vrieze, D. E. Weeks, B. S. Weir, S. T. Weiss, L.-C. Weng, C. J. Willer, Y. Zhang, X. Zhao, D. K. Arnett, A. E. Ashley-Koch, K. C. Barnes, E. Boerwinkle, S. Gabriel, R. Gibbs, K. M. Rice, S. S. Rich, E. K. Silverman, P. Qasba, W. Gan, NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, G. J. Papanicolaou, D. A. Nickerson, S. R. Browning, M. C. Zody, S. Zöllner, J. G. Wilson, L. A. Cupples, C. C. Laurie, C. E. Jaquish, R. D. Hernandez, T. D. O'Connor, G. R. Abecasis, Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature*. **590**, 290–299 (2021).

28. S. Shi, N. Yuan, M. Yang, Z. Du, J. Wang, X. Sheng, J. Wu, J. Xiao, Comprehensive Assessment of Genotype Imputation Performance. *Hum. Hered.* **83**, 107–116 (2018).

29. V. Tam, N. Patel, M. Turcotte, Y. Bossé, G. Paré, D. Meyre, Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.* **20**, 467–484 (2019).

30. R. J. Klein, C. Zeiss, E. Y. Chew, J.-Y. Tsai, R. S. Sackler, C. Haynes, A. K. Henning, J. P. SanGiovanni, S. M. Mane, S. T. Mayne, M. B. Bracken, F. L. Ferris, J. Ott, C. Barnstable, J. Hoh, Complement factor H polymorphism in age-related macular degeneration. *Science*. **308**, 385–389 (2005).

31. D. F. Easton, K. A. Pooley, A. M. Dunning, P. D. P. Pharoah, D. Thompson, D. G. Ballinger, J. P. Struewing, J. Morrison, H. Field, R. Luben, N. Wareham, S. Ahmed, C. S. Healey, R. Bowman, SEARCH collaborators, K. B. Meyer, C. A. Haiman, L. K. Kolonel, B. E. Henderson, L. Le Marchand, P. Brennan, S. Sangrajrang, V. Gaborieau, F. Odefrey, C.-Y. Shen, P.-E. Wu, H.-C. Wang, D. Eccles, D. G. Evans, J. Peto, O. Fletcher, N. Johnson, S. Seal, M. R. Stratton, N. Rahman, G. Chenevix-Trench, S. E. Bojesen, B. G. Nordestgaard, C. K. Axelsson, M. Garcia-Closas, L. Brinton, S. Chanock, J. Lissowska, B. Peplonska, H. Nevanlinna, R. Fagerholm, H. Eerola, D. Kang, K.-Y. Yoo, D.-Y. Noh, S.-H. Ahn, D. J. Hunter, S. E. Hankinson, D. G. Cox, P. Hall, S. Wedren, J. Liu, Y.-L. Low, N. Bogdanova, P. Schürmann, T. Dörk, R. A. E. M. Tollenaar, C. E. Jacobi, P. Devilee, J. G. M. Klijn, A. J. Sigurdson, M. M. Doody, B. H. Alexander, J. Zhang, A. Cox, I. W. Brock, G. MacPherson, M. W. R. Reed, F. J. Couch, E. L. Goode, J. E. Olson, H. Meijers-Heijboer, A. van den Ouweland, A. Uitterlinden, F. Rivadeneira, R. L. Milne, G. Ribas, A. Gonzalez-Neira, J. Benitez, J. L. Hopper, M. McCredie, M. Southey, G. G. Giles, C. Schroen, C. Justenhoven, H. Brauch, U. Hamann, Y.-D. Ko, A. B. Spurdle, J. Beesley, X. Chen, kConFab, AOCS Management Group, A. Mannermaa, V.-M. Kosma, V. Kataja, J. Hartikainen, N. E. Day, D. R. Cox, B. A. J. Ponder, Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature*. **447**, 1087–1093 (2007).

32. G. Thomas, K. B. Jacobs, P. Kraft, M. Yeager, S. Wacholder, D. G. Cox, S. E. Hankinson, A. Hutchinson, Z. Wang, K. Yu, N. Chatterjee, M. Garcia-Closas, J. Gonzalez-Bosquet, L. Prokunina-Olsson, N. Orr, W. C. Willett, G. A. Colditz, R. G. Ziegler, C. D. Berg, S. S. Buys, C. A. McCarty, H. S. Feigelson, E. E. Calle, M. J. Thun, R. Diver, R. Prentice, R. Jackson, C. Kooperberg, R. Chlebowski, J. Lissowska, B. Peplonska, L. A. Brinton, A. Sigurdson, M. Doody, P. Bhatti, B. H. Alexander, J. Buring, I.-M. Lee, L. J. Vatten, K. Hveem, M. Kumle, R. B. Hayes, M. Tucker, D. S. Gerhard, J. F. Fraumeni Jr, R. N. Hoover, S. J. Chanock, D. J. Hunter, A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (RAD51L1). *Nat. Genet.* **41**, 579–584 (2009).

33. S. N. Stacey, A. Manolescu, P. Sulem, T. Rafnar, J. Gudmundsson, S. A. Gudjonsson, G. Masson, M. Jakobsdottir, S. Thorlacius, A. Helgason, K. K. Aben, L. J. Strobbe, M. T. Albers-Akkers, D. W. Swinkels, B. E. Henderson, L. N. Kolonel, L. Le Marchand, E. Millastre, R. Andres, J. Godino, M. D. Garcia-Prats, E. Polo, A. Tres, M. Mouy, J. Saemundsdottir, V. M. Backman, L. Gudmundsson, K. Kristjansson, J. T. Bergthorsson, J. Kostic, M. L. Frigge, F. Geller, D. Gudbjartsson, H. Sigurdsson, T. Jonsdottir, J. Hrafnkelsson, J. Johannsson, T. Sveinsson, G. Myrdal, H. N. Grimsson, T. Jonsson, S. von Holst, B. Werelius, S. Margolin, A. Lindblom, J. I. Mayordomo, C. A. Haiman, L. A. Kiemeney, O. T. Johannsson, J. R. Gulcher, U. Thorsteinsdottir, A. Kong, K. Stefansson, Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. *Nat. Genet.* **39**, 865–869 (2007).

34. S. Ahmed, G. Thomas, M. Ghoussaini, C. S. Healey, M. K. Humphreys, R. Platte, J. Morrison, M. Maranian, K. A. Pooley, R. Luben, D. Eccles, D. G. Evans, O. Fletcher, N. Johnson, I. dos Santos Silva, J. Peto, M. R. Stratton, N. Rahman, K. Jacobs, R. Prentice, G. L. Anderson, A. Rajkovic, J. D. Curb, R. G. Ziegler, C. D. Berg, S. S. Buys, C. A. McCarty, H. S. Feigelson, E. E. Calle, M. J. Thun, W. R. Diver, S. Bojesen, B. G. Nordestgaard, H. Flyger, T. Dörk, P. Schürmann, P. Hillemanns, J. H. Karstens, N. V. Bogdanova, N. N. Antonenkova, I. V. Zalutsky, M. Bermisheva, S. Fedorova, E. Khusnutdinova, SEARCH, D. Kang, K.-Y. Yoo, D. Y. Noh, S.-H. Ahn, P. Devilee, C. J. van Asperen, R. A. E. M. Tollenaar, C. Seynaeve, M. Garcia-Closas, J. Lissowska, L. Brinton, B. Peplonska, H. Nevanlinna, T. Heikkinen, K. Aittomäki, C. Blomqvist, J. L. Hopper, M. C. Southey, L. Smith, A. B. Spurdle, M. K. Schmidt, A. Broeks, R. R. van Hien, S. Cornelissen, R. L. Milne, G. Ribas, A. González-Neira, J. Benitez, R. K. Schmutzler, B. Burwinkel, C. R. Bartram, A. Meindl, H. Brauch, C. Justenhoven, U. Hamann, GENICA Consortium, J. Chang-Claude, R. Hein, S. Wang-Gohrke, A. Lindblom, S. Margolin, A. Mannermaa, V.-M. Kosma, V. Kataja, J. E. Olson, X. Wang, Z. Fredericksen, G. G. Giles, G. Severi, L. Baglietto, D. R. English, S. E. Hankinson, D. G. Cox, P. Kraft, L. J. Vatten, K. Hveem, M. Kumle, A. Sigurdson, M. Doody, P. Bhatti, B. H. Alexander, M. J. Hooning, A. M. W. van den Ouweland, R. A. Oldenburg, M. Schutte, P. Hall, K. Czene, J. Liu, Y. Li, A. Cox, G. Elliott, I. Brock, M. W. R. Reed, C.-Y. Shen, J.-C. Yu, G.-C. Hsu, S.-T. Chen, H. Anton-Culver, A. Ziogas, I. L. Andrulis, J. A. Knight, kConFab, Australian Ovarian Cancer Study Group, J. Beesley, E. L. Goode, F. Couch, G. Chenevix-Trench, R. N. Hoover, B. A. J. Ponder, D. J. Hunter, P. D. P. Pharoah, A. M. Dunning, S. J. Chanock, D. F. Easton, Newly discovered breast cancer susceptibility loci on 3p24 and 17q23.2. *Nat. Genet.* **41**, 585–590 (2009).

35. B. Gold, T. Kirchhoff, S. Stefanov, J. Lautenberger, A. Viale, J. Garber, E. Friedman, S.

Narod, A. B. Olshen, P. Gregersen, K. Kosarin, A. Olsh, J. Bergeron, N. A. Ellis, R. J. Klein, A. G. Clark, L. Norton, M. Dean, J. Boyd, K. Offit, Genome-wide association study provides evidence for a breast cancer risk locus at 6q22.33. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 4340–4345 (2008).

36. W. Zheng, J. Long, Y.-T. Gao, C. Li, Y. Zheng, Y.-B. Xiang, W. Wen, S. Levy, S. L. Deming, J. L. Haines, K. Gu, A. M. Fair, Q. Cai, W. Lu, X.-O. Shu, Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1. *Nat. Genet.* **41**, 324–328 (2009).

37. D. J. Hunter, P. Kraft, K. B. Jacobs, D. G. Cox, M. Yeager, S. E. Hankinson, S. Wacholder, Z. Wang, R. Welch, A. Hutchinson, J. Wang, K. Yu, N. Chatterjee, N. Orr, W. C. Willett, G. A. Colditz, R. G. Ziegler, C. D. Berg, S. S. Buys, C. A. McCarty, H. S. Feigelson, E. E. Calle, M. J. Thun, R. B. Hayes, M. Tucker, D. S. Gerhard, J. F. Fraumeni Jr, R. N. Hoover, G. Thomas, S. J. Chanock, A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat. Genet.* **39**, 870–874 (2007).

38. J. Gudmundsson, P. Sulem, T. Rafnar, J. T. Bergthorsson, A. Manolescu, D. Gudbjartsson, B. A. Agnarsson, A. Sigurdsson, K. R. Benediktsdottir, T. Blondal, M. Jakobsdottir, S. N. Stacey, J. Kostic, K. T. Kristinsson, B. Birgisdottir, S. Ghosh, D. N. Magnusdottir, S. Thorlacius, G. Thorleifsson, S. L. Zheng, J. Sun, B.-L. Chang, J. B. Elmore, J. P. Breyer, K. M. McReynolds, K. M. Bradley, B. L. Yaspan, F. Wiklund, P. Stattin, S. Lindström, H.-O. Adami, S. K. McDonnell, D. J. Schaid, J. M. Cunningham, L. Wang, J. R. Cerhan, J. L. St Sauver, S. D. Isaacs, K. E. Wiley, A. W. Partin, P. C. Walsh, S. Polo, M. Ruiz-Echarri, S. Navarrete, F. Fuertes, B. Saez, J. Godino, P. C. Weijerman, D. W. Swinkels, K. K. Aben, J. A. Witjes, B. K. Suarez, B. T. Helfand, M. L. Frigge, K. Kristjansson, C. Ober, E. Jonsson, G. V. Einarsson, J. Xu, H. Gronberg, J. R. Smith, S. N. Thibodeau, W. B. Isaacs, W. J. Catalona, J. I. Mayordomo, L. A. Kiemeney, R. B. Barkardottir, J. R. Gulcher, U. Thorsteinsdottir, A. Kong, K. Stefansson, Common sequence variants on 2p15 and Xp11.22 confer susceptibility to prostate cancer. *Nat. Genet.* **40**, 281–283 (2008).

39. R. A. Eeles, Z. Kote-Jarai, A. A. Al Olama, G. G. Giles, M. Guy, G. Severi, K. Muir, J. L. Hopper, B. E. Henderson, C. A. Haiman, J. Schleutker, F. C. Hamdy, D. E. Neal, J. L. Donovan, J. L. Stanford, E. A. Ostrander, S. A. Ingles, E. M. John, S. N. Thibodeau, D. Schaid, J. Y. Park, A. Spurdle, J. Clements, J. L. Dickinson, C. Maier, W. Vogel, T. Dörk, T. R. Rebbeck, K. A. Cooney, L. Cannon-Albright, P. O. Chappuis, P. Hutter, M. Zeegers, R. Kaneva, H.-W. Zhang, Y.-J. Lu, W. D. Foulkes, D. R. English, D. A. Leongamornlert, M. Tymrakiewicz, J. Morrison, A. T. Ardern-Jones, A. L. Hall, L. T. O'Brien, R. A. Wilkinson, E. J. Saunders, E. C. Page, E. J. Sawyer, S. M. Edwards, D. P. Dearnaley, A. Horwich, R. A. Huddart, V. S. Khoo, C. C. Parker, N. Van As, C. J. Woodhouse, A. Thompson, T. Christmas, C. Ogden, C. S. Cooper, M. C. Southey, A. Lophatananon, J.-F. Liu, L. N. Kolonel, L. Le Marchand, T. Wahlfors, T. L. Tammela, A. Auvinen, S. J. Lewis, A. Cox, L. M. FitzGerald, J. S. Koopmeiners, D. M. Karyadi, E. M. Kwon, M. C. Stern, R. Corral, A. D. Joshi, A. Shahabi, S. K. McDonnell, T. A. Sellers, J. Pow-Sang, S. Chambers, J. Aitken, R. A. F. Gardiner, J. Batra, M. A. Kedda, F. Lose, A. Polanowski, B. Patterson, J. Serth, A. Meyer, M. Luedeke, K. Stefflova, A. M. Ray, E. M. Lange, J. Farnham, H. Khan, C. Slavov, A. Mitkova, G. Cao, UK Genetic Prostate Cancer Study Collaborators/British Association of

Urological Surgeons' Section of Oncology, UK ProtecT Study Collaborators, PRACTICAL Consortium, D. F. Easton, Identification of seven new prostate cancer susceptibility loci through a genome-wide association study. *Nat. Genet.* **41**, 1116–1121 (2009).

40. J. Gudmundsson, P. Sulem, D. F. Gudbjartsson, T. Blondal, A. Gylfason, B. A. Agnarsson, K. R. Benediktsdottir, D. N. Magnusdottir, G. Orlygsdottir, M. Jakobsdottir, S. N. Stacey, A. Sigurdsson, T. Wahlfors, T. Tammela, J. P. Breyer, K. M. McReynolds, K. M. Bradley, B. Saez, J. Godino, S. Navarrete, F. Fuertes, L. Murillo, E. Polo, K. K. Aben, I. M. van Oort, B. K. Suarez, B. T. Helfand, D. Kan, C. Zanon, M. L. Frigge, K. Kristjansson, J. R. Gulcher, G. V. Einarsson, E. Jonsson, W. J. Catalona, J. I. Mayordomo, L. A. Kiemeney, J. R. Smith, J. Schleutker, R. B. Barkardottir, A. Kong, U. Thorsteinsdottir, T. Rafnar, K. Stefansson, Genome-wide association and replication studies identify four variants associated with prostate cancer susceptibility. *Nat. Genet.* **41**, 1122–1126 (2009).

41. R. A. Eeles, Z. Kote-Jarai, G. G. Giles, A. A. A. Olama, M. Guy, S. K. Jugurnauth, S. Mulholland, D. A. Leongamornlert, S. M. Edwards, J. Morrison, H. I. Field, M. C. Southey, G. Severi, J. L. Donovan, F. C. Hamdy, D. P. Dearnaley, K. R. Muir, C. Smith, M. Bagnato, A. T. Ardern-Jones, A. L. Hall, L. T. O'Brien, B. N. Gehr-Swain, R. A. Wilkinson, A. Cox, S. Lewis, P. M. Brown, S. G. Jhavar, M. Tymrakiewicz, A. Lophatananon, S. L. Bryant, UK Genetic Prostate Cancer Study Collaborators, British Association of Urological Surgeons' Section of Oncology, UK ProtecT Study Collaborators, A. Horwich, R. A. Huddart, V. S. Khoo, C. C. Parker, C. J. Woodhouse, A. Thompson, T. Christmas, C. Ogden, C. Fisher, C. Jamieson, C. S. Cooper, D. R. English, J. L. Hopper, D. E. Neal, D. F. Easton, Multiple newly identified loci associated with prostate cancer susceptibility. *Nat. Genet.* **40**, 316–321 (2008).

42. G. Thomas, K. B. Jacobs, M. Yeager, P. Kraft, S. Wacholder, N. Orr, K. Yu, N. Chatterjee, R. Welch, A. Hutchinson, A. Crenshaw, G. Cancel-Tassin, B. J. Staats, Z. Wang, J. Gonzalez-Bosquet, J. Fang, X. Deng, S. I. Berndt, E. E. Calle, H. S. Feigelson, M. J. Thun, C. Rodriguez, D. Albanes, J. Virtamo, S. Weinstein, F. R. Schumacher, E. Giovannucci, W. C. Willett, O. Cussenot, A. Valeri, G. L. Andriole, E. D. Crawford, M. Tucker, D. S. Gerhard, J. F. Fraumeni Jr, R. Hoover, R. B. Hayes, D. J. Hunter, S. J. Chanock, Multiple loci identified in a genome-wide association study of prostate cancer. *Nat. Genet.* **40**, 310–315 (2008).

43. J. Gudmundsson, P. Sulem, A. Manolescu, L. T. Amundadottir, D. Gudbjartsson, A. Helgason, T. Rafnar, J. T. Bergthorsson, B. A. Agnarsson, A. Baker, A. Sigurdsson, K. R. Benediktsdottir, M. Jakobsdottir, J. Xu, T. Blondal, J. Kostic, J. Sun, S. Ghosh, S. N. Stacey, M. Mouy, J. Saemundsdottir, V. M. Backman, K. Kristjansson, A. Tres, A. W. Partin, M. T. Albers-Akkers, J. Godino-Ivan Marcos, P. C. Walsh, D. W. Swinkels, S. Navarrete, S. D. Isaacs, K. K. Aben, T. Graif, J. Cashy, M. Ruiz-Echarri, K. E. Wiley, B. K. Suarez, J. A. Witjes, M. Frigge, C. Ober, E. Jonsson, G. V. Einarsson, J. I. Mayordomo, L. A. Kiemeney, W. B. Isaacs, W. J. Catalona, R. B. Barkardottir, J. R. Gulcher, U. Thorsteinsdottir, A. Kong, K. Stefansson, Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. *Nat. Genet.* **39**, 631–637 (2007).

44. L. T. Amundadottir, P. Sulem, J. Gudmundsson, A. Helgason, A. Baker, B. A. Agnarsson, A. Sigurdsson, K. R. Benediktsdottir, J.-B. Cazier, J. Sainz, M. Jakobsdottir, J. Kostic, D. N. Magnusdottir, S. Ghosh, K. Agnarsson, B. Birgisdottir, L. Le Roux, A. Olafsdottir, T.

Blondal, M. Andresdottir, O. S. Gretarsdottir, J. T. Bergthorsson, D. Gudbjartsson, A. Gylfason, G. Thorleifsson, A. Manolescu, K. Kristjansson, G. Geirsson, H. Isaksson, J. Douglas, J.-E. Johansson, K. Bälter, F. Wiklund, J. E. Montie, X. Yu, B. K. Suarez, C. Ober, K. A. Cooney, H. Gronberg, W. J. Catalona, G. V. Einarsson, R. B. Barkardottir, J. R. Gulcher, A. Kong, U. Thorsteinsdottir, K. Stefansson, A common variant associated with prostate cancer in European and African populations. *Nat. Genet.* **38**, 652–658 (2006).

45. M. L. Freedman, C. A. Haiman, N. Patterson, G. J. McDonald, A. Tandon, A. Waliszewska, K. Penney, R. G. Steen, K. Ardlie, E. M. John, I. Oakley-Girvan, A. S. Whittemore, K. A. Cooney, S. A. Ingles, D. Altshuler, B. E. Henderson, D. Reich, Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 14068–14073 (2006).

46. M. Yeager, N. Orr, R. B. Hayes, K. B. Jacobs, P. Kraft, S. Wacholder, M. J. Minichiello, P. Fearnhead, K. Yu, N. Chatterjee, Z. Wang, R. Welch, B. J. Staats, E. E. Calle, H. S. Feigelson, M. J. Thun, C. Rodriguez, D. Albanes, J. Virtamo, S. Weinstein, F. R. Schumacher, E. Giovannucci, W. C. Willett, G. Cancel-Tassin, O. Cussenot, A. Valeri, G. L. Andriole, E. P. Gelmann, M. Tucker, D. S. Gerhard, J. F. Fraumeni Jr, R. Hoover, D. J. Hunter, S. J. Chanock, G. Thomas, Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat. Genet.* **39**, 645–649 (2007).

47. A. A. Al Olama, Z. Kote-Jarai, G. G. Giles, M. Guy, J. Morrison, G. Severi, D. A. Leongamornlert, M. Tymrakiewicz, S. Jhavar, E. Saunders, J. L. Hopper, M. C. Southey, K. R. Muir, D. R. English, D. P. Dearnaley, A. T. Ardern-Jones, A. L. Hall, L. T. O'Brien, R. A. Wilkinson, E. Sawyer, A. Lophatananon, UK Genetic Prostate Cancer Study Collaborators/British Association of Urological Surgeons' Section of Oncology, UK Prostate testing for cancer and Treatment study (ProtecT Study) Collaborators, A. Horwich, R. A. Huddart, V. S. Khoo, C. C. Parker, C. J. Woodhouse, A. Thompson, T. Christmas, C. Ogden, C. Cooper, J. L. Donovan, F. C. Hamdy, D. E. Neal, R. A. Eeles, D. F. Easton, Multiple loci on 8q24 associated with prostate cancer susceptibility. *Nat. Genet.* **41**, 1058–1060 (2009).

48. M. Yeager, N. Chatterjee, J. Ciampa, K. B. Jacobs, J. Gonzalez-Bosquet, R. B. Hayes, P. Kraft, S. Wacholder, N. Orr, S. Berndt, K. Yu, A. Hutchinson, Z. Wang, L. Amundadottir, H. S. Feigelson, M. J. Thun, W. R. Diver, D. Albanes, J. Virtamo, S. Weinstein, F. R. Schumacher, G. Cancel-Tassin, O. Cussenot, A. Valeri, G. L. Andriole, E. D. Crawford, C. A. Haiman, B. Henderson, L. Kolonel, L. Le Marchand, A. Siddiq, E. Riboli, T. J. Key, R. Kaaks, W. Isaacs, S. Isaacs, K. E. Wiley, H. Gronberg, F. Wiklund, P. Stattin, J. Xu, S. L. Zheng, J. Sun, L. J. Vatten, K. Hveem, M. Kumle, M. Tucker, D. S. Gerhard, R. N. Hoover, J. F. Fraumeni Jr, D. J. Hunter, G. Thomas, S. J. Chanock, Identification of a new prostate cancer susceptibility locus on chromosome 8q24. *Nat. Genet.* **41**, 1055–1057 (2009).

49. J. Gudmundsson, P. Sulem, V. Steinthorsdottir, J. T. Bergthorsson, G. Thorleifsson, A. Manolescu, T. Rafnar, D. Gudbjartsson, B. A. Agnarsson, A. Baker, A. Sigurdsson, K. R. Benediktsdottir, M. Jakobsdottir, T. Blondal, S. N. Stacey, A. Helgason, S. Gunnarsdottir, A. Olafsdottir, K. T. Kristinsson, B. Birgisdottir, S. Ghosh, S. Thorlacius, D. Magnusdottir, G. Stefansdottir, K. Kristjansson, Y. Bagger, R. L. Wilensky, M. P. Reilly, A. D. Morris, C. H. Kimber, A. Adeyemo, Y. Chen, J. Zhou, W.-Y. So, P. C. Y. Tong, M. C. Y. Ng, T. Hansen,

G. Andersen, K. Borch-Johnsen, T. Jorgensen, A. Tres, F. Fuertes, M. Ruiz-Echarri, L. Asin, B. Saez, E. van Boven, S. Klaver, D. W. Swinkels, K. K. Aben, T. Graif, J. Cashy, B. K. Suarez, O. van Vierssen Trip, M. L. Frigge, C. Ober, M. H. Hofker, C. Wijmenga, C. Christiansen, D. J. Rader, C. N. A. Palmer, C. Rotimi, J. C. N. Chan, O. Pedersen, G. Sigurdsson, R. Benediktsson, E. Jonsson, G. V. Einarsson, J. I. Mayordomo, W. J. Catalona, L. A. Kiemeney, R. B. Barkardottir, J. R. Gulcher, U. Thorsteinsdottir, A. Kong, K. Stefansson, Two variants on chromosome 17 confer prostate cancer risk, and the one in TCF2 protects against type 2 diabetes. *Nat. Genet.* **39**, 977–983 (2007).

50. J. Sun, S. L. Zheng, F. Wiklund, S. D. Isaacs, G. Li, K. E. Wiley, S.-T. Kim, Y. Zhu, Z. Zhang, F.-C. Hsu, A. R. Turner, P. Stattin, W. Liu, J. W. Kim, D. Duggan, J. Carpten, W. Isaacs, H. Grönberg, J. Xu, B.-L. Chang, Sequence variants at 22q13 are associated with prostate cancer risk. *Cancer Res.* **69**, 10–15 (2009).

51. Z. K. Stadler, P. Thom, M. E. Robson, J. N. Weitzel, N. D. Kauff, K. E. Hurley, V. Devlin, B. Gold, R. J. Klein, K. Offit, Genome-wide association studies of cancer. *J. Clin. Oncol.* **28**, 4255–4267 (2010).

52. Implications for immunosurveillance of altered HLA class I phenotypes in human tumours. *Immunol. Today*. **18**, 89–95 (1997).

53. D. J. Hicklin, F. M. Marincola, S. Ferrone, HLA class I antigen downregulation in human cancers: T-cell immunotherapy revives an old story. *Mol. Med. Today*. **5**, 178–186 (1999).

54. A. K. Johnsen, D. J. Templeton, M. Sy, C. V. Harding, Deficiency of transporter for antigen presentation (TAP) in tumor cells allows evasion of immune surveillance and increases tumorigenesis. *J. Immunol.* **163**, 4224–4231 (1999).

55. N. P. Restifo, F. Esquivel, Y. Kawakami, J. W. Yewdell, J. J. Mulé, S. A. Rosenberg, J. R. Bennink, Identification of human cancers deficient in antigen processing. *J. Exp. Med.* **177**, 265–272 (1993).

56. R. Rotem-Yehudar, M. Groettrup, A. Soza, P. M. Kloetzel, R. Ehrlich, LMP-associated proteolytic activities and TAP-dependent peptide transport for class 1 MHC molecules are suppressed in cell lines transformed by the highly oncogenic adenovirus 12. *J. Exp. Med.* **183**, 499–514 (1996).

57. M. H. Lind, B. Rozell, R. P. A. Wallin, M. van Hogerlinden, H.-G. Ljunggren, R. Toftgård, I. Sur, Tumor necrosis factor receptor 1-mediated signaling is required for skin cancer development induced by NF-kappaB inhibition. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 4972–4977 (2004).

58. E. Y. Lin, V. Gouon-Evans, A. V. Nguyen, J. W. Pollard, The macrophage growth factor CSF-1 in mammary gland development and tumor progression. *J. Mammary Gland Biol. Neoplasia*. **7**, 147–162 (2002).

59. E. M. Sotomayor, Y. X. Fu, M. Lopez-Cepero, L. Herbert, J. J. Jimenez, C. Albarracin, D. M. Lopez, Role of tumor-derived cytokines on the immune system of mice bearing a

mammary adenocarcinoma. II. Down-regulation of macrophage-mediated cytotoxicity by tumor-derived granulocyte-macrophage colony-stimulating factor. *J. Immunol.* **147**, 2816–2823 (1991).

60. J. F. Grosso, M. N. Jure-Kunkel, CTLA-4 blockade in tumor models: an overview of preclinical and translational research. *Cancer Immun.* **13**, 5 (2013).

61. J. C. Yang, M. Hughes, U. Kammula, R. Royal, R. M. Sherry, S. L. Topalian, K. B. Suri, C. Levy, T. Allen, S. Mavroukakis, I. Lowy, D. E. White, S. A. Rosenberg, Ipilimumab (anti-CTLA4 antibody) causes regression of metastatic renal cell cancer associated with enteritis and hypophysitis. *J. Immunother.* **30**, 825–830 (2007).

62. T. J. Lynch, I. Bondarenko, A. Luft, P. Serwatowski, F. Barlesi, R. Chacko, M. Sebastian, J. Neal, H. Lu, J.-M. Cuillerot, M. Reck, Ipilimumab in combination with paclitaxel and carboplatin as first-line treatment in stage IIIB/IV non-small-cell lung cancer: results from a randomized, double-blind, multicenter phase II study. *J. Clin. Oncol.* **30**, 2046–2054 (2012).

63. S. I. Grivennikov, F. R. Greten, M. Karin, Immunity, inflammation, and cancer. *Cell*. **140**, 883–899 (2010).

64. G. P. Dunn, A. T. Bruce, H. Ikeda, L. J. Old, R. D. Schreiber, Cancer immunoediting: from immunosurveillance to tumor escape. *Nat. Immunol.* **3**, 991–998 (2002).

65. M. Mangino, M. Roederer, M. H. Beddall, F. O. Nestle, T. D. Spector, Innate and adaptive immune traits are differentially affected by genetic and environmental factors. *Nat. Commun.* **8**, 13850 (2017).

66. M. F. Keller, A. P. Reiner, Y. Okada, F. J. A. van Rooij, A. D. Johnson, M.-H. Chen, A. V. Smith, A. P. Morris, T. Tanaka, L. Ferrucci, A. B. Zonderman, G. Lettre, T. Harris, M. Garcia, S. Bandinelli, R. Qayyum, L. R. Yanek, D. M. Becker, L. C. Becker, C. Kooperberg, B. Keating, J. Reis, H. Tang, E. Boerwinkle, Y. Kamatani, K. Matsuda, N. Kamatani, Y. Nakamura, M. Kubo, S. Liu, A. Dehghan, J. F. Felix, A. Hofman, A. G. Uitterlinden, C. M. van Duijn, O. H. Franco, D. L. Longo, A. B. Singleton, B. M. Psaty, M. K. Evans, L. A. Cupples, J. I. Rotter, C. J. O'Donnell, A. Takahashi, J. G. Wilson, S. K. Ganesh, M. A. Nalls, CHARGE Hematology, COGENT, BioBank Japan Project (RIKEN) Working Groups, Trans-ethnic meta-analysis of white blood cell phenotypes. *Hum. Mol. Genet.* **23**, 6944–6960 (2014).

67. V. Orrù, M. Steri, G. Sole, C. Sidore, F. Virdis, M. Dei, S. Lai, M. Zoledziewska, F. Busonero, A. Mulas, M. Floris, W. I. Mentzen, S. A. M. Urru, S. Olla, M. Marongiu, M. G. Piras, M. Lobina, A. Maschio, M. Pitzalis, M. F. Urru, M. Marcelli, R. Cusano, F. Deidda, V. Serra, M. Oppo, R. Pilu, F. Reinier, R. Berutti, L. Pireddu, I. Zara, E. Porcu, A. Kwong, C. Brennan, B. Tarrier, R. Lyons, H. M. Kang, S. Uzzau, R. Atzeni, M. Valentini, D. Firinu, L. Leoni, G. Rotta, S. Naitza, A. Angius, M. Congia, M. B. Whalen, C. M. Jones, D. Schlessinger, G. R. Abecasis, E. Fiorillo, S. Sanna, F. Cucca, Genetic variants regulating immune cell levels in health and disease. *Cell*. **155**, 242–256 (2013).

68. J. Ye, K. M. Gillespie, S. Rodriguez, Unravelling the Roles of Susceptibility Loci for

Autoimmune Diseases in the Post-GWAS Era. *Genes* . **9** (2018), doi:10.3390/genes9080377.

69. R. W. Sayaman, M. Saad, V. Thorsson, W. Hendrickx, J. Roelands, Y. Mokrab, F. Farshidfar, T. Kirchhoff, R. F. Sweis, O. F. Bathe, E. Porta-Pardo, M. J. Campbell, C. Stretch, D. Hu, S. Huntsman, R. E. Graff, N. Syed, L. Radvanyi, S. Shelley, D. Wolf, F. M. Marincola, M. Ceccarelli, J. Galon, E. Ziv, D. Bedognetti, Germline genetic contribution to the immune landscape of cancer, , doi:10.1101/2020.01.30.926527.

70. Impact of Genetic Polymorphisms on Human Immune Cell Gene Expression. *Cell*. **175**, 1701–1715.e16 (2018).

71. J. Zhou, C. L. Theesfeld, K. Yao, K. M. Chen, A. K. Wong, O. G. Troyanskaya, Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat. Genet.* **50**, 1171–1179 (2018).

72. N. Mavaddat, K. Michailidou, J. Dennis, M. Lush, L. Fachal, A. Lee, J. P. Tyrer, T.-H. Chen, Q. Wang, M. K. Bolla, X. Yang, M. A. Adank, T. Ahearn, K. Aittomäki, J. Allen, I. L. Andrulis, H. Anton-Culver, N. N. Antonenkova, V. Arndt, K. J. Aronson, P. L. Auer, P. Auvinen, M. Barrdahl, L. E. Beane Freeman, M. W. Beckmann, S. Behrens, J. Benitez, M. Bermisheva, L. Bernstein, C. Blomqvist, N. V. Bogdanova, S. E. Bojesen, B. Bonanni, A.-L. Børresen-Dale, H. Brauch, M. Bremer, H. Brenner, A. Brentnall, I. W. Brock, A. Brooks-Wilson, S. Y. Brucker, T. Brüning, B. Burwinkel, D. Campa, B. D. Carter, J. E. Castelao, S. J. Chanock, R. Chlebowski, H. Christiansen, C. L. Clarke, J. M. Collée, E. Cordina-Duverger, S. Cornelissen, F. J. Couch, A. Cox, S. S. Cross, K. Czene, M. B. Daly, P. Devilee, T. Dörk, I. Dos-Santos-Silva, M. Dumont, L. Durcan, M. Dwek, D. M. Eccles, A. B. Ekici, A. H. Eliassen, C. Ellberg, C. Engel, M. Eriksson, D. G. Evans, P. A. Fasching, J. Figueroa, O. Fletcher, H. Flyger, A. Försti, L. Fritschi, M. Gabrielson, M. Gago-Dominguez, S. M. Gapstur, J. A. García-Sáenz, M. M. Gaudet, V. Georgoulias, G. G. Giles, I. R. Gilyazova, G. Glendon, M. S. Goldberg, D. E. Goldgar, A. González-Neira, G. I. Grenaker Alnæs, M. Grip, J. Gronwald, A. Grundy, P. Guénel, L. Haeberle, E. Hahnen, C. A. Haiman, N. Håkansson, U. Hamann, S. E. Hankinson, E. F. Harkness, S. N. Hart, W. He, A. Hein, J. Heyworth, P. Hillemanns, A. Hollestelle, M. J. Hooning, R. N. Hoover, J. L. Hopper, A. Howell, G. Huang, K. Humphreys, D. J. Hunter, M. Jakimovska, A. Jakubowska, W. Janni, E. M. John, N. Johnson, M. E. Jones, A. Jukkola-Vuorinen, A. Jung, R. Kaaks, K. Kaczmarek, V. Kataja, R. Keeman, M. J. Kerin, E. Khusnutdinova, J. I. Kiiski, J. A. Knight, Y.-D. Ko, V.-M. Kosma, S. Koutros, V. N. Kristensen, U. Krüger, T. Kühl, D. Lambrechts, L. Le Marchand, E. Lee, F. Lejbkowicz, J. Lilyquist, A. Lindblom, S. Lindström, J. Lissowska, W.-Y. Lo, S. Loibl, J. Long, J. Lubiński, M. P. Lux, R. J. MacInnis, T. Maishman, E. Makalic, I. Maleva Kostovska, A. Mannermaa, S. Manoukian, S. Margolin, J. W. M. Martens, M. E. Martinez, D. Mavroudis, C. McLean, A. Meindl, U. Menon, P. Middha, N. Miller, F. Moreno, A. M. Mulligan, C. Mulot, V. M. Muñoz-Garzon, S. L. Neuhausen, H. Nevanlinna, P. Neven, W. G. Newman, S. F. Nielsen, B. G. Nordestgaard, A. Norman, K. Offit, J. E. Olson, H. Olsson, N. Orr, V. S. Pankratz, T.-W. Park-Simon, J. I. A. Perez, C. Pérez-Barrios, P. Peterlongo, J. Peto, M. Pinchev, D. Plaseska-Karanfilska, E. C. Polley, R. Prentice, N. Presneau, D. Prokofyeva, K. Purrington, K. Pylkäs, B. Rack, P. Radice, R. Rau-Murthy, G. Rennert, H. S. Rennert, V. Rhenius, M. Robson, A. Romero, K. J. Ruddy, M. Ruebner, E. Saloustros, D. P. Sandler, E. J. Sawyer, D. F. Schmidt, R. K. Schmutzler, A. Schneeweiss, M. J. Schoemaker, F.

Schumacher, P. Schürmann, L. Schwentner, C. Scott, R. J. Scott, C. Seynaeve, M. Shah, M. E. Sherman, M. J. Shrubsole, X.-O. Shu, S. Slager, A. Smeets, C. Sohn, P. Soucy, M. C. Southey, J. J. Spinelli, C. Stegmaier, J. Stone, A. J. Swerdlow, R. M. Tamimi, W. J. Tapper, J. A. Taylor, M. B. Terry, K. Thöne, R. A. E. M. Tollenaar, I. Tomlinson, T. Truong, M. Tzardi, H.-U. Ulmer, M. Untch, C. M. Vachon, E. M. van Veen, J. Vijai, C. R. Weinberg, C. Wendt, A. S. Whittemore, H. Wildiers, W. Willett, R. Winqvist, A. Wolk, X. R. Yang, D. Yannoukakos, Y. Zhang, W. Zheng, A. Ziogas, ABCTB Investigators, kConFab/AOCS Investigators, NBCS Collaborators, A. M. Dunning, D. J. Thompson, G. Chenevix-Trench, J. Chang-Claude, M. K. Schmidt, P. Hall, R. L. Milne, P. D. P. Pharoah, A. C. Antoniou, N. Chatterjee, P. Kraft, M. García-Closas, J. Simard, D. F. Easton, Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. *Am. J. Hum. Genet.* **104**, 21–34 (2019).

73. C. M. Lewis, E. Vassos, Polygenic risk scores: from research tools to clinical instruments. *Genome Med.* **12**, 44 (2020).

74. J. Morales, D. Welter, E. H. Bowler, M. Cerezo, L. W. Harris, A. C. McMahon, P. Hall, H. A. Junkins, A. Milano, E. Hastings, C. Malangone, A. Buniello, T. Burdett, P. Flicek, H. Parkinson, F. Cunningham, L. A. Hindorff, J. A. L. MacArthur, A standardized framework for representation of ancestry data in genomics studies, with application to the NHGRI-EBI GWAS Catalog. *Genome Biology*. **19** (2018), , doi:10.1186/s13059-018-1396-2.

CHAPTER 1: Germline modifiers of the tumor immune microenvironment implicate drivers of cancer risk and immunotherapy response

1.1.1 Abstract

With the continued promise of immunotherapy as an avenue for treating cancer, understanding how host genetics contributes to the tumor immune microenvironment (TIME) is essential to tailoring cancer screening and treatment strategies. Approaches that intersect SNP modifiers of molecular phenotype, such as gene expression, with disease phenotypes have shown promise for implicating causal genetic factors. Here we evaluated 194 literature-curated TIME associations and 890 associations detected with 157 immune phenotype (IP) components found using genotypes from over 8,000 individuals in The Cancer Genome Atlas. Of these 1084, 233 associations comprising 219 unique TIME-SNPs were also cancer relevant, associating with cancer risk, survival, and/or immunotherapy treatment response. Many cancer relevant TIME-SNPS overlapped regions of active transcription, and were associated with gene expression in specific immune cell subsets, such as macrophages and dendritic cells. TIME-SNPs associated with cancer risk and response to immunotherapy implicated genes involved in antigen presentation, especially by antigen presenting cells. The strongest associations with survival were with *PD-L1* and *CTLA-4*, suggesting that SNPs modifying the potential for immune evasion could contribute to disease progression. To assess whether our approach could reveal novel cancer immunotherapy targets, we inhibited *CTSS,* a gene implicated by cancer risk and immunotherapy response-associated TIME-SNPs; CTSS inhibition resulted in slowed tumor growth and extended survival *in vivo*. These results validate the potential of cancer relevant TIME-SNPs to implicate target genes for countering immune suppressive characteristics of the TIME and set the stage for future host genetics analysis integrating germline variation and TIME characteristics.

1.1.2 Introduction

Cancer is a disease characterized by heterogeneous somatic and germline mutations that promote abnormal cellular growth, evasion from the immune system, dysregulation of cellular energetics, and inflammation(*1–4*). Both inflammation and immune surveillance contribute to the selective forces that shape tumor evolution(*3–6*). Immunotherapies alleviating immune suppressive signals have emerged as a promising treatment strategy; however, response rates are low and the determinants of response remain elusive(*7*, *8*). Furthermore, the potential of galvanizing the immune system is still unmet due to an incomplete understanding of the complex tumor immune microenvironment (TIME). In particular, knowledge of germline factors and other intrinsic factors that interact with characteristics of tumors to render them sensitive to host-immunity or immunotherapy is lacking.

Efforts to identify germline variation associated with anti-tumor immune responses have pointed to effects on immune infiltration levels and immune pathways, such as *TGF-β* and *IFN-γ(9–11)*. Genes with significant cis-eQTLs in the TCGA are both enriched for immune-related genes and associated with immune cell abundance within the TIME(*12*). These studies provide evidence that variants may act through specific effects on immune cells. eQTL profiling of 15 sorted immune cell subsets from healthy individuals found that the effects of many eQTLs were specific to immune cell subsets(*13*). Understanding mechanisms and cell-type effects of TIME host genetic interactions could not only identify aspects of immunity that negatively impact cancer and immunotherapy outcomes, but also point to putative targetable cell types and molecules for modulating immune responses.

Here, we sought to identify common germline variants associated with TIME characteristics that are also associated with cancer outcomes, reasoning that such dual associations

would implicate the aspects of immunity most critical for tumor control and uncover putative new targets for immunotherapy(*14*, *15*). We combined a new analysis of The Cancer Genome Atlas (TCGA) with previous germline studies to identify 1084 TIME associations and further analyzed them to converge on a subset associated with cancer outcomes. Adaptive immune genes relating to antigen processing and presentation, including several class I and II MHC genes, were implicated in risk for multiple cancers in the UK Biobank and validated in independent cohorts. We also found immune checkpoint and Th17 variants underlying immune evasion to be associated with overall and progression-free survival, respectively. Lastly, we identified 11 SNPs associating with ICB response in publicly available cohorts. Since overlapping eQTL and GWAS signals can suggest causal associations, we selected one such gene, *CTSS,* for validation as a possible immunotherapy target and found that inhibition of CTSS prolonged survival and reduced tumor growth in an MC38 murine model. These results illuminate the role of common genetic variation underlying the TIME in cancer risk and survival while also providing a potential novel avenue for immunotherapy target discovery. The study design is summarized in **Figure 1.1.**

1.1.3 Results

1.1.3.1. Identifying Heritable Characteristics of the Tumor Immune Microenvironment (TIME)

To focus on common germline genetics with the potential to modify tumor immune responses, we assessed which characteristics of the TIME showed evidence of SNP heritability. To describe the TIME, we collected a comprehensive set of immune phenotype ("IP") components comprising composite measures derived from bulk gene expression and expression levels of individual immune-related genes (**Figure 1.1**). Composite phenotypes included infiltrating

immune cell levels calculated using CIBERSORTx (immune infiltrates) and 6 immune subtype scores from a pan-cancer TCGA analysis by Thorsson et al. (landscape components).

Immunomodulators were collected from Thorsson et al., where weighted gene correlation network analysis was used as an unbiased systematic approach to identify gene sets relevant to the TIME. We included genes from these sets along with immune checkpoint genes, cell type markers, antigen presentation genes, TGF-β pathway genes, and IFN-γ genes as these have been implicated as important modifiers of the TIME**.** After removing IP components with high numbers of zero values to reduce spurious associations, we retained 724 immune-related genes and 9 composite phenotypes (733 IP components total) measured across 30 cancer types **(Table 1.1, Figure S1.1)**.

We evaluated the potential of germline variation to explain inter-tumor differences in IP components by performing SNP heritability analysis **(Figure 1.1)**. Since highly polymorphic regions such as the HLA locus can inflate SNP heritability estimates, we separately estimated SNP heritability attributable to the HLA locus and the rest of the genome. We identified 235 (32.0%) IP components where levels were SNP-heritable (>5% of variance in expression or composite value was attributable to genetic variance; *i.e.* Vg/Vp > 5%; **Figure 1.1**). For these 235 IP components, we conducted 2-state GCTA analysis and identified 140 (59.6%) that had a significant proportion of SNP heritability attributable to regions outside the HLA locus, while 17 (7.2%) were mostly attributable to the HLA locus at an FDR < 0.05. We focused our TIME-SNP discovery analysis on these 157 SNP-heritable IP components.

1.1.3.2 Detecting putative germline modifiers of the tumor immune microenvironment

To build a comprehensive dataset of germline variants associated with the TIME, we integrated associations with SNP-heritable IP components and immune SNPs previously implicated by published studies. To identify genetic variants underlying IP component SNP

heritability, we performed a genome-wide association study (GWAS). First, we performed GWAS for each of the 140 heritable IP components outside of the HLA locus across individuals of European ancestry in the TCGA **(Figure S1.2)**. Only common germline variants with minor allele frequency > 1% were considered and imputation quality (Rsq) was evaluated to ensure high accuracy (**Figure S1.2)**. No evidence of inflation was observed (**Figure S1.2).** Using linkage and distance-based clumping(*16*), we identified 825 associations with 75 SNP-heritable IP components at a threshold of 7.1×10^{-8} (Bonferroni-corrected suggestive threshold of 1×10^{-5}), with 545 of the 825 (66.1%) passing a Bonferroni-corrected genome-wide significance threshold of 3.6×10^{-10} (**Figure 1.2**). These 825 TIME associations implicated 795 unique TIME-SNPs. *Cis* associations, defined as an associated locus occurring within 1 MB of an IP component gene transcription start site, encompassed the majority (95.0%) of associations(*17*), while 5.0% of the associations were *trans*. Mechanisms of *trans* associations are complex and tend to have weaker effects on transcriptional regulation(*18*). In contrast, *cis* associations are proximal to an IP component and have more direct effects on transcription. Overall, *ERAP2* (181, 21.9%), *CCBL2* (76, 9.2%), *DHFR* (75 9.0%) and *ERAP1* (70, 8.5%) had the most germline associations **(Figure S1.2)** of the 140 IP components tested.

To remove HLA region associations solely attributable to LD structure(*19*, *20*), we conducted conditional GWAS analysis for seventeen IP components corresponding to genes in the HLA region of chromosome 6, iteratively adding the strongest associated SNP on chromosome 6 as a covariate until no significant SNPs remained. Alignment to a general HLA gene reference can introduce error into expression level estimates due to the highly polymorphic nature of these genes. We therefore also revisited SNP associations with gene expression estimates derived from allele-specific RNA alignments(*21*) and performed GWAS analysis using allele specific expression. In total, 63 independent TIME-SNPs and 65 TIME associations were associated with HLA region

gene expression **(Figure 1.2)**; MHC Class II genes, *HLA-DQB1* and *HLA-DRB5,* had 8 and 7 significant LD-independent associations, respectively. Generally, LD-independent SNPs clustered by genomic regions with *HLA-A, HLA-B, HLA-C* associated variants falling in the MHC Class I genomic region and *HLA-DQB1, HLA-DQA1, HLA-DPB1, HLA-DRB5* associated variants falling in the MHC Class II genomic region **(Figure S1.2)**. rs17612852 was associated with *HLA-DQB1* expression using both the reference alignment and allele-specific expression. Combining GWAS and conditional HLA GWAS associations, we identified 890 TIME associations and 858 unique TIME-SNPs.

We noted some correlation among IP components across tumors, especially for components associated with macrophages and lymphocytes which were among the most abundant infiltrating immune cells **(Figure S1.2)**. We investigated whether IP component correlation would inflate the chance of detecting SNPs associated with a particular group, however analysis of summary statistics showed that despite their correlation, IP components typically did not recover the same SNP associations unless the IP components were genes encoded at the same genomic locus, such as *ERAP1* and *LNPEP* and *OAS1* and *OAS3* (**Figure S1.3)**. Clustering of pearson correlation values amongst SNP-heritable IP components revealed 2 major groups of genes (**Figure S1.3)**. The largest group included MHC Class I and II genes along with macrophage genes *VSIG4, CD163, FCGR2A FCGR3A, HAVCR2, LILRB2, LILRB4* and *CD53* **(Figure S1.3)** and was most strongly associated with antigen presentation, dendritic cell processing, and *IL-10* production (**Figure S1.3).** The next largest group comprised two anti-correlated subgroups of genes which contained *EP300* and *TREX1* respectively **(Figure S1.3)**. This group of genes were related to innate immune activation, the C-type lectin receptor signaling pathway and antigen presentation **(Figure S1.3)**. Two major groups of genes correlated with the top 2 principal components from Principal Component Analysis (PCA) conducted on the expression of the 157 unique SNP-

heritable IP components across TCGA tumors. *CD53, CD86* and *CYBB*, which are highly correlated ($\rho > 0.7$) to the Thorsson et al.(*22*) Macrophage Regulation score, were major contributors to PC1 while *HACD2, LNPEP* and *EP300,* were major contributors to PC2.

Previous studies of germline variation and important modulators of immune checkpoint response, such as APOE(*23*), CTSW(*24*), CTLA-4(*25*), PD-L1(*26*, *27*), PD-1(*28–30*), CXCR3/CCR5(*31*), IRF5(*32*) and FGFR4(*33*) along with immune signatures and immune cell infiltration have been conducted(*10*, *12*, *34*). We incorporated these 194 germline associations from literature into our analyses **(Figure 1.2)**. Like Shahamatdar et al.(*10*), we included immune infiltrates estimated from bulk RNA sequencing into the set of immune components we investigated, however, our filter on the proportion of phenotypic variance explained by genotype eliminated one of the two associations reported by Shahamatdar et al. which related to T follicular helper cell infiltration (V(g)/V(p) < 0.000001 whereas we required V(g)/V(p) >= 0.05). Zhang et al.(*24*) took a fundamentally different approach, analyzing ER+ breast cancer-associated variants from Michailidou et al.(*35*) for proximity to immunoinflammatory GWAS SNPs. The top SNP, rs3903072, was an eQTL for *CTSW* in breast cancer. Although not specifically focussed on breast cancer, our study also identified *CTSW* as a SNP-heritable IP component (GCTA V(g)/V(p)= 12.1%) and detected a pan-cancer association with rs3903072 (beta = 0.21, p=2.8e-36). The study by Sayaman et al.(*34*) focused on 139 immune traits described in the Thorsson et al.(*22*) paper, of which 106 were immune signatures and 33 included immune measures such as TCR/BCR characteristics, CIBERSORTx infiltration and antigen load. Comparing gene results between Sayaman et al. and our study, 10 genes were shared between our analyses, *HLA-DRB5, HLA-B, HLA-DRB1, MICB, HLA-DQB1, HLA-DQB2, HLA-DQA1, HLA-DQA2, MICA, HLA-C*, emphasizing the importance of MHC Class I and II machinery in modifying the TIME. Of our variants, 31 were in linkage disequilibrium (LD) ($R^2 > 0.20$) with 485 of 598 Sayaman et al. snps.

Using $R^2 > 0.50$, 19 of our variants were in LD with 361 Sayaman et al. SNPs.

Combining TIME-associations and literature SNPs resulted in a set of 1084 candidate SNPs. A number of TIME-SNPs were associated with multiple IP components, thus we had a greater number of associations than TIME-SNPs. For example, within our own discovery pipeline, rs2693076 was associated with *LILRB2, PLEK, MYO1F,* and *CD14*. From literature curation, Sayaman et al. identified associations with rs2111485 and multiple signatures, including interferon-signaling and *IFIT3* signaling.

1.1.3.3 Identification and Characterization of TIME-SNPs Related to Cancer Outcomes.

A screen to detect TIME-SNPs would be expected to detect variants associated with immune traits more generally, including variants that may have little relevance to cancer. To focus our analysis on TIME-SNPs with evidence for cancer relevance, we evaluated their association with disease risk, progression and response to immunotherapy. SNPs associated with any of these aspects were considered cancer relevant.

SNPs were designated cancer risk associated if they had previously been implicated by a cancer GWAS study in the NHGRI-EBI GWAS catalog(*36*, *37*), Vanderbilt PheWAS catalog(*38*) or were found to associate with an ICD10 code related to cancer in a PheWAS of the UK Biobank(*39*). Rather than a GWAS which analyzes many genetic variants with one phenotype, a PheWAS analyzes many phenotypes compared to a single genetic variant. In our study, we conducted a PheWAS in the UK Biobank with TIME-SNPs and identified 61 cancer risk associations, We also identified 64 cancer risk associations through intersection with the NHGRI-EBI GWAS catalog and 92 associations through intersection with the Vanderbilt PheWAS catalog. Although we used different methods to assess cancer risk, we observed high overlap in risk variants identified by the three sources **(Figure S1.4)**. When assessing overlap based on the corresponding

IP components, a higher degree of overlap was observed, with only 2 SNP-heritable IP components, *TAP2* and *LNPEP,* being uniquely implicated by the UK Biobank **(Figure S1.4)**. In total, 138 unique TIME-SNPs associated with 41 IP components had cancer risk associations. This included risk associations with non-melanoma skin cancer, melanoma, lung cancer, prostate cancer, breast cancer and head and neck squamous cell carcinoma. Most cancer risk associations were found with the MHC II gene signature described in Sayaman *et al*.(*40*) (31), *CTSS* expression (12) and *ERAP2* expression (11).

We next evaluated association of TIME-SNPs with overall and progression-free survival in the TCGA. We identified 92 associations and 87 variants that were significantly associated with overall or progression-free survival in at least one tumor type (FDR < 0.05). Of the variants associated with overall survival, the majority of associations occurred in cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC, 8), uterine carcinosarcoma (UCS, 8) and thyroid cancer (THCA, 7). Variants associated with progression-free survival were also found in cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC, 6), but otherwise were most frequent in stomach adenocarcinoma (STAD, 5), rectum adenocarcinoma (READ, 5), and liver hepatocellular carcinoma (LIHC, 5). Variants associated with *PD-1*, the MHC II signature, *GPLD1, ERAP1, ERAP2, CTSW*, dendritic cell signatures, *VAMP3* and *FAM216A* were associated with survival in more than 1 tumor type in the TCGA. For example, the intronic *PD-L1* variant, rs822339, was associated with overall survival in multiple cancer types, including lung adenocarcinoma (LUAD), kidney chromophobe cancer (KICH) and thyroid cancer (THCA), confirming survival associations reported by Yoshida *et al(26)*. Of these 87 variants, 5 remained significantly associated with survival in a Cox proportional-hazards analysis with covariates, including 3 *PD-L1* variants, an *ERAP2* variant and a *DCTN5* variant.

To investigate the implication of TIME-SNPs for immune checkpoint blockade (ICB)

response, we collected sequencing and ICB response information for 276 patients with melanoma treated with immune checkpoint inhibitors from 4 studies(*41–45*), and imputed SNPs from exome sequencing data. Accuracy of exome-based imputation was assessed by comparing original TCGA genotype calls to genotypes imputed in from TCGA exome data at positions matching those in the ICB data; aside from variants on chromosome 6 within the HLA region most variants were accurately imputed**.** Ultimately, of the 1084 TIME-SNPs, we considered 525 that could be imputed with sufficient quality (minor allele frequency > 0.05 in all 4 discovery ICB cohorts with imputation accuracy of at least 0.3(*46*, *47*). For ICB variant discovery, we conducted meta-analysis with METAL(*48*) and found that 6 SNPs were significantly associated with ICB response (FDR < 0.25), implicating IP components such as *PSMD11, ERAP1, TREX1, ERAP2* and a T follicular helper cell signature. With a less stringent FDR threshold (FDR < 0.5), 11 variants were associated with ICB response, implicating the following additional genes: *CTSS, FAM216A, DHFR, DCTN5, LYZ.*

In total, these analyses implicated 219 TIME-SNPs and 233 associations as cancer relevant. Assessing overlap based on SNPs passing significance thresholds, 12 variants associated with *DBNDD1, CCBL2, ERAP2, GPLD1, DCTN5, HLA-DQA1, HLA-DRB5, HLA-DQA1, CTLA-4* and the MHC II signature were implicated in both risk and survival analysis. The *CTSS* variant, rs2305814, was associated with cancer risk and ICB response, the *DHFR* variant, rs434130, was associated with cancer survival and ICB response and the *DCTN5* variant, rs546055 was associated with cancer risk, survival and ICB response. IP components *CTSS, ERAP1, TREX1, ERAP2* and *DCTN5* were implicated in both cancer risk and immunotherapy response while *FAM216A, ERAP1, TREX1, ERAP2, DHFR, DCTN5,* and *LYZ* were implicated in both cancer prognosis and immunotherapy response.

Focusing on these 219 cancer relevant TIME-SNPs, we sought to understand what aspects

of the tumor-immune interface were affected. Cancer relevant TIME-SNPs were associated with 65 unique IP components. Several of these were antigen presentation and macrophage regulation genes, including both MHC I and II pathway genes **(Figure 1.3)**. Fifty-two of the 233 (22.3%) associations were from literature curation and included associations with the MHC II signature, the T follicular helper cell signature, IFN, the IFIT3 attractor signature, *CTLA-4*, *PD-1*, *PD-L1*, *APOE*, *CTSW*, and monocyte, dendritic cell and TH2 cell infiltration. Of the 181 (77.7%) associations from our TIME germline discovery pipeline, the majority of variants were detected as *cis* associations (96.1%), aside from 7 (3.9%) *trans* associations (**Figure S1.4).** Ten cancer relevant TIME-SNPs (4.6%) affected protein-coding regions **(Figure S1.4).** In the case of *CTSS, HLA-B, TAP2, HLA-A, CTLA-4* and *APOE*, missense variants in coding regions were associated with expression differences. In addition, missense variants in *AP5B1* and *ERAP1* were associated with expression differences in *CTSW* and *ERAP2,* respectively **(Figure S1.4)** and missense variants in *EGFL8* and *CCHCR1* were associated with differences in the MHC II signature described by Sayaman et al.

As the majority of TIME-SNPs fell within non-coding genomic regions, we evaluated their potential to affect regulatory sites. Histone marks provide important information about regulation of chromatin architecture and accessibility of DNA sequence for transcription(*49*). Regions harboring TIME-SNPs were strongly enriched in H3K27ac, H3K36me3 and H3K4me3 histone marks and depleted in H3K9me3 histone marks **(Figure 1.3)** (*50*). H3K27ac is a known marker of active enhancers and H3K4me3 is usually enriched at promoters near transcription start sites(*51*, *52*) suggesting some TIME-SNPs are broadly associated with transcription while others may be gene specific. Coincidentally, TIME-SNPs were depleted in repressive H3K9me3 marks(*53*). In addition to overall enrichment in markers of active transcription, enrichment in particular histone marks was more pronounced in certain immune cell types. The strongest enrichment in *cis*-

regulatory elements was seen in PBMCs, CD14+ monocytes, B cells and CD8+ T cells.

To obtain more information about TIME-SNP cell-type specific effects, we evaluated whether TIME-SNP association with gene expression in TCGA was dependent on immune cell infiltration level or corresponded to known immune eQTLs in DICE. Of the 219 variants, 37 were macrophage cell-type eQTLs, 32 were CD4+ T cell-type eQTLs, 26 were CD8+ T cell-type eQTLs and 22 were B cell-type eQTLs **(Figure 1.3)**. Comparing myeloid-specific eQTLs to lymphoid-specific eQTLs, variants associated with *TAP2, CTLA-4, FCGR3B, ERAP2* and *DBNDD1* were myeloid-specific. Looking specifically at cell types implicated by risk, survival and immunotherapy response, macrophages, CD4+ and CD8+ T cells were the cell types with the most associations **(Figures S1.4)**. These findings reveal a subset of TIME-SNPs that specifically modify the activity of genes in immune rather than tumor cells and indicate cell types that may contribute to cancer risk, progression and immunotherapy response.

1.1.3.5 TIME-SNPs underlying antigen presentation stratify melanoma and prostate cancer risk

We further analyzed genes with TIME-SNP cancer risk associations in the UK Biobank. At a strict FDR threshold (FDR < 0.05), 57 TIME-SNPs were associated with cancer risk, 28 of which were associated with genes critical for antigen presentation, such as *CTSS, ERAP1, CTSW, ERAP2* and MHC Class I and II genes. Using a less stringent FDR threshold (FDR < 0.20), we identified 119 TIME-SNPs associated with cancer risk, including 65 variants associated with genes critical for antigen presentation.

We noted an EP300 variant weakly associated with risk of neoplasms of digestive organs (FDR < 17%). *EP300* can potentiate MHC Class I antigen presentation(*54*) and *EP300* somatic mutations have been reported in multiple types of cancer. Downregulation of *EP300* has been

associated with higher anti-tumor immunity and suggesting that *EP300* inhibition could potentiate immune checkpoint blockade(*55*, *56*). However, in work by Kruper et al., higher EP300 expression was associated with more rapid tumor cell proliferation and sensitivity to ICB, while lower activity conferred resistance, suggesting instead that EP300 loss may potentiate immune evasion(*54*). In our case, the variant allele was associated both with increased expression in the TCGA and with higher cancer risk in the UK Biobank, suggesting that more rapid growth could outweigh the benefit of immune evasion for early digestive tumors(*55*). Cell-type interaction analysis did not reveal *EP300* variant interaction with any specific immune cell types.

To further validate antigen presentation TIME-SNPs as *bona fide* cancer risk SNPs, we asked whether a polygenic risk score (PRS) using antigen presentation TIME-SNPs would generalize to independent cancer cohorts. A TIME-SNP melanoma PRS restricted to associations with antigen presentation genes and the MHC Class II signature described in Sayaman et al. stratified patients with high vs low risk in the UK Biobank **(Figure 1.4)** and validated in an independent cohort of 3029 melanoma cases and controls for UT MD Anderson(*57*). Although the difference in PRS score distributions for cases and controls was small **(Figure 1.4),** the odds of melanoma were significantly different in the top and bottom 10th quantile in the validation cohort **(Figure 1.4).** The MHC II Signature described in Sayaman et al. was strongly correlated with MHC class II genes, such as *HLA-DQB1,* but also weakly correlated with MHC Class I gene expression **(Figure S1.5).**

Eleven variants were associated both with prostate cancer risk in the UK Biobank (FDR < 0.2) and with the MHC II signature or antigen presentation genes, with 1 SNP in common with the melanoma PRS. A prostate cancer TIME-SNP PRS constructed from these in the UKBB stratified patients with high vs low risk **(Figure 1.4)** and validated in an independent cohort of 91,644 prostate cancer cases and controls for ELLIPSE Consortium(*58*). Although the difference in PRS

score distributions for cases and controls was small in the validation cohort **(Figure 1.4),** the odds of prostate cancer were significantly different in the top and bottom 10th quantile **(Figure 1.4).**

Notably, 14 of the melanoma PRS snps and 4 of the prostate cancer PRS snps were DICE eQTLs. Of the melanoma PRS snps, 10 were macrophage cell-type eQTLs and 10 were CD8+ T cell-type eQTLs. This suggested that the PRS might be related to cancer risk through modulation of the inflammatory landscape. Indeed, in the TCGA, tumors in the upper 10th quantile of the melanoma TIME-SNP PRS had higher levels of infiltration by pro-tumor inflammatory M2-like **(Figure 1.4),** but not M0 or M1-like macrophages. Promotion of an inflammatory pro-tumor environment was also correlated with decreased CD8+ T cell infiltration **(Figure 1.4).**

1.1.3.6 Variants underlying immune evasion are associated with cancer survival

We revisited the 92 significant associations and 87 unique variants implicated via Kaplan-Meier Analysis (FDR < 0.05) using the burden of SNPs in each of the categories shown in Figure 3A by performing a within tumor type Cox Proportional-Hazards analysis with covariates including age of diagnosis, gender, and stage of cancer, requiring associations to pass an FDR < 0.05. We found a significant association of immune checkpoint variant burden with overall survival in lung adenocarcinoma (FDR < 0.05) (**Figure 1.5A, S1.6).** In the case of Th17 signature, Th2 signature, and dendritic cell signatures, we only had 1 variant per category. Nonetheless, these variants were significantly associated with progression-free survival (FDR < 0.05) (**Figure 1.5, S1.6)**. The variant associated with Th17 was an *IL17RA* intronic variant. We observed worse survival (**Figure 1.5)** in liver and hepatocellular carcinoma patients with this variant that increased *IL17RA* expression in TCGA (**Figure 1.5).** Indeed, this confirms reports that IL-17 is associated with worse survival in hepatocellular carcinoma through the potential pro-inflammatory role of IL-17(*59*, *60*).

We also found the burden of immune checkpoint variants to be significantly associated with overall survival in lung adenocarcinoma (**Figure 1.5).** Three variants in the burden score were associated with *PD-L1* and 2 with *CTLA-4*. Two *PD-L1* variants were in regions associated with the gene promoter while 1 was in a region associated with H3K27ac marks, suggesting variants may modify transcription of the genes **(Figure S1.6).** Interestingly, while we did not see correlation of the burden score with *PD-L1* expression in bulk tumor RNA, we observed weak correlation with expression of *PD-1* and *CTLA-4* **(Figure 1.5)**. Individuals with mid-level burden had increased expression of immune checkpoints and worse survival compared to individuals with a lower burden of variants, consistent with the immuno-inhibitory roles of these molecules. However, individuals with high immune checkpoint burden had significantly decreased immune checkpoint expression. This observation could point to an adaptive response of the tumor immune microenvironment or even potential non-genetic compensatory mechanisms that have developed to avoid high immune checkpoint expression throughout life. We investigated correlation of the burden score with tumor mutation burden (TMB) in lung cancers but found no significant correlation **(Figure S1.6)**. Since immune checkpoint molecules have very cell-type specific expression, we investigated correlation of the burden score with expression of *PD-L1* specifically in immune cell types (**Figure 1.5)** and indeed found positive correlation in immune cell subsets, except for stimulated CD8+ and CD4+ T cells. These results confirm that TIME-SNPs have cell-type effects, as we observed stronger effects of TIME-SNPs on immune cell specific expression compared to bulk RNA expression in the TCGA.

1.1.3.7 TIME-SNPs Implicate Targets for Modulating Immune Responses

To evaluate whether cancer relevant TIME SNPs implicate aspects of tumor immunity that could serve as an entry points for immunotherapy, we further analyzed the 11 variants associated

with ICB response, and their associated IP components **(Figure 1.6)**. We confirmed that TIME-SNP genotypes were similarly distributed and free of artifacts across all cohorts by PCA analysis **(Figure S1.7).** The direction of effect of variants associated with responder status was mostly consistent across cohorts, though there were some differences observed with the *PSMD11* variant, rs28459155 (**Figure 1.6)**. This variant was associated with lower odds of being a responder in Miao et al. and Hugo et al. but higher odds of being a responder in Van Allen et al., Snyder et al. and Riaz et al. As a comparison to current ICB biomarkers, we also evaluated association of tumor mutation burden (TMB) and expression levels of *PD-L1, PD-1,* and *CTLA-4* with responder status and found no significant associations **(Figure 1.6)**. We ran associations with the 11 variants and *TMB, PD-L1, PD-1,* and *CTLA-4* to determine if any variants were associated with these previously researched biomarkers. We observed an association between TMB and *ERAP1* variant rs27765 in Hugo et al. (**Figure S1.7).**

We constructed a burden score with the response-associated allele for all 11 variants across the four melanoma ICB cohorts and confirmed that ICB responders had a significantly higher score compared to non-responders (**Figure 1.6).** The association between higher burden score and ICB response remained significant in two additional independent validation cohorts from Miao et al. (renal cell carcinoma) and Rizvi et al. (non-small cell lung cancer) **(Figure 1.6)**. We repeated this analysis selecting 11 TIME-SNPs at random, matched for minor allele frequency, and found that the observed difference in burden score between responders and nonresponders was significantly larger than random in both discovery and validation sets **(Figure S1.7).** Although we had a small sample size, we built a polygenic risk score trained on the discovery melanoma cohort using PRSice and were able to achieve an AUC of at least 0.6 in Rizvi et al. and Miao et al. validation sets **(Figure S1.7).** We then evaluated how much of the variation in responder status was explained by germline burden score, TMB, TMB plus immune checkpoint markers, and all of these markers

together. Germline burden score explained some variance in responder status in all 6 cohorts (2.5-25.4%). In Van Allen et al., Hugo et al., and Riaz et al., germline burden score explained more variance than TMB and immune checkpoint markers. In Van Allen et al. and Riaz et al., the combination of all markers maximized the variance explained **(Figure 1.6)**.

Colocalization of gene expression and GWAS signals can point to putative causal disease-related genes. To determine if genes implicated by TIME-SNPs could reveal candidate immunotherapy targets, we analyzed the genes associated with the 11 variants. These genes once again emphasized a central role for antigen processing and presentation pathways (GO:0048002 antigen processing and presentation of peptide antigen [fold enrichment = 90.33, FDR < 0.0351], GO:0030333 antigen processing and presentation [fold enrichment = 82.46, FDR < 0.0011]). Examining gene expression available for 4 of the 6 cohorts, we noted that none of these genes were significantly differentially expressed between ICB responders and nonresponders**.** However, some TIME-SNPs were associated with high expression of an IP component gene and worse outcome, uncovering potential targets to validate. Two *TREX1* variants were associated with immunotherapy response; rs11917071 was associated with increased *TREX1* expression in Van Allen et al. and lower odds of being a responder (**Figure S1.7),** and rs2267844 was associated with decreased *TREX1* expression in Hugo et al. and Miao et al. and higher odds of being a responder **(Figure S1.7)**. These results are consistent with findings that *TREX1* acts as an immunoinhibitor that prevents cGAS-STRING initiation, with inhibition of *TREX1* stimulating IFN$\boldsymbol{\gamma}$ signaling and autoimmunity, making it a potential immunomodulatory target(*61*, *62*). *PSMD11* variant rs28459155 was associated with lower odds of being a responder but increased expression of *PSMD11* **(Figure S1.7)**. *PSMD11*, a proteosomal protein involved in ubiquitination, is associated with worse prognosis in pancreatic cancer(*63*). Similarly, *CTSS* variant rs23058814 was associated with higher odds of being a responder and decreased *CTSS* expression in Van Allen et al. **(Figure**

S1.7). Increased *CTSS* expression has been linked to tumor progression in follicular lymphoma due to decreased CD8+ T cell recruitment(*64*). In two separate mouse models, significant differences in *Ctss* expression was observed between immunotherapy responders and non-responders (**Figure S1.7).** Furthermore, we observed increased M1 macrophage infiltration in individuals with *CTSS* variant in Hugo et al. **(Figure S1.7).**

Based on these association data, we hypothesized the inhibition of *CTSS* would improve ICB response. To test this hypothesis, we treated mice implanted with MC38 tumors with a CTSS small molecule inhibitor. Mice treated with CTSS inhibitor had slowed tumor growth and better survival compared to control mice (**Figure 1.6).** We also evaluated the interaction of CTSS inhibitor treatment with anti-PD-1. Mice treated with CTSS inhibitor or anti-PD-1 monotherapy had significantly decreased tumor growth and better survival compared to control mice. Additionally, tumor growth was further decreased in mice treated with the combination of anti-PD-1 and CTSS inhibitor as compared to mice treated with anti-PD-1 or CTSS inhibitor alone. In the MC38 model, we observed an increase in infiltrating M1 macrophages and a decrease in M2 macrophages similar to findings from Hugo et al. (**Figure 1.6).** These findings demonstrate that a focused screen for cancer relevant TIME-associated variants provides a fruitful strategy to reveal novel immunotherapy targets. Furthermore, the influence of *CTSS* inhibition on the myeloid landscape, identifies macrophages as potential cell type that may modulate immunotherapy response.

1.1.4 Discussion

The success of immunotherapies has generated enthusiasm for using the human immune system as a weapon to eliminate cancers(*65–68*). However the very existence of cancer indicates the failure of the immune system to control malignant cell populations throughout multiple stages

of tumor development(*4*). Here we studied genetic variants associated with interindividual differences in immune traits and the tumor immune microenvironment, reasoning that these variants could reveal the aspects of immunity most critical for the successful immune control of tumors. We focused on a subset of immune characteristics that showed evidence of SNP heritability in The Cancer Genome Atlas or were implicated in the literature and that were also associated with cancer outcomes including risk, survival and response to immune checkpoint inhibitors. Out of the 1084 SNP associations with 129 immune phenotype (IP) components, ultimately 219 TIME-SNPs associating with 65 IP components met these criteria. 104 of these SNPs were not in LD ($R^2 < 0.5$) with genome-wide significant SNPs reported by cancer GWAS studies or Vanderbilt PheWAS catalog despite interacting with important clinical outcomes. In addition, we demonstrated that our approach could implicate putative targets to modify anti-tumor immunity; *TREX1* has previously been highlighted as a promising target(*61*), and small molecule inhibition of *CTSS* resulted in slower tumor growth and longer survival of mice, with effects comparable to anti-PD-1. While several studies suggest that CTSS may have immune suppressive roles in follicular lymphoma(*64*, *69*, *70*), this is the first time to the best of our knowledge that inhibition of this gene was shown to relieve immune suppression in solid tumors.

The immune system interacts with tumors throughout their development and treatment, both through tumor-promoting inflammation and immune-mediated elimination of cancerous cells(*71*, *72*). Cancer relevant SNP associations implicated IP components related to antigen presentation (MHC Class I and II, *CTSS*, *ERAP1*) and inflammation (Th17, Th2). Limited sample sizes and the inability to impute HLA region SNPs with sufficient quality in exome-only cohorts impeded the complete assessment of pleiotropy in the context of cancer relevance. Nonetheless, we found several SNPs linked to genes involved in antigen presentation and evasion of the adaptive

immune response associated with multiple aspects of the tumor-immune interplay. More generally, we saw that IP components implicated in cancer risk were mainly those involved in both MHC Class I and Class II antigen presentation, while SNPs associated with prognosis pointed to genes that would support evasion of the MHC I CD8+ T cell axis including *PD-L1, CTLA-4* and Th17.

Adaptive immunity is crucial for the host anti-tumor immune responses. In cancer, downregulation of antigen presentation machinery is a frequent mechanism of immune evasion. MHC Class I and II genotypes shape mutational landscapes in cancer and inform immunotherapy response(*73–75*). Interestingly, we found multiple germline variants associated with both MHC Class I and Class II genes were also associated with cancer risk. This included several antigen presentation pathway genes not directly encoding the MHC and located outside of the HLA region: *CTSS, CTSW, ERAP1, ERAP2,* and *TAP2. ERAP1* and *ERAP2* are endoplasmic reticulum peptidases that trim peptides before loading them onto MHC proteins(*76*, *77*). *ERAP1/ERAP2* polymorphisms have been associated with cervical cancer and autoimmunity(*78–84*). *CTSS* is a cysteine protease critical for MHC Class II loading and is frequently mutated in follicular lymphoma. Its loss limits communication with CD4+ T follicular helper cells while inducing antigen diversification and activation of CD8+ T cells(*64*, *69*). *CTSW* is crucial for cytotoxicity and is expressed in specific immune cell types(*69*). Interestingly, the involvement of MHC II and immune cell specific genes suggest that inter-individual variation in immune surveillance contributes to cancer risk. In further support of this, polygenic risk scores constructed from MHC II pathway associated germline variants were able to stratify melanoma and prostate cancer risk in multiple cohorts. These two tumor types fall at opposite ends of the spectrum of immune activity, with melanoma being one of the most immunotherapy responsive tumors, and prostate cancer being one of the least(*85*, *86*). Furthermore, MHC Class II expression is linked to ICB response

in melanoma(*87*). Although prostate cancer is considered immunologically "cold", rare dramatic responses to immunotherapy have been documented(*88*). MHC Class II is usually restricted to professional antigen presenting cells although prostate cancer cells have been shown to express MHC Class II. The Class II pathway is crucial for a prolonged anti-tumor response as it leads to sustained CD8+ T cell activation and leads to more complete tumor clearance. Due to low imputation quality, we were not able to assess the dependence of ICB responses on HLA region SNPs, however *CTSS* was both detected and validated as a determinant of response, suggesting MHC Class II could also underlie response to immunotherapy. Together with reports from multiple immune vaccines studies that responses were primarily driven by CD4+ T cells(*89–92*), these findings place further emphasis on the central importance of MHC II for effective anti-tumor immune responses.

Innate immunity is the branch of the immune system that acts as the body's first line of defense against microbial pathogens and cancer cells, and involves cells originating in the bone marrow that carry non-polymorphic receptors. Cells of the innate branch of the immune system such as macrophages and dendritic cells play a pivotal role in the tumor microenvironment creating a hostile pro-inflammatory environment, suppressing T cells, promoting angiogenesis, and initiating lymphangiogenesis. *TREX1*, *TLR2*, *VAMP3* and *APOE* were among innate immune genes implicated by our screen. In a recent study it was shown that mice with one missense mutation (D18N) in the TREX1 gene were protected from tumor growth, and that protection was associated with a reduced expression of *PD-1* in T lymphocytes (aka less exhaustion)(*93*). This study supports the view that a missense mutation of TREX1 leads to enhanced tumor T cell immunity. Using RNA-seq analysis, we found individuals with *TREX1*-reducing variants had higher odds of being ICB responders. Arguably, TREX1 inhibition could be regarded as a viable

anticancer immunotherapeutic strategy.

In general, the innate compartment was implicated mostly through cell type eQTL analysis. Variants associated with *ERAP1, CTSW, DCTN5, MICA, BTN3A2* and MHC Class I and II were cell for cells in both innate and adaptive subsets. *CTSW* was a cell-type eQTL for macrophages and dendritic cells, as well as CD8+ T cells and B cells. Although *CTSW* has strong immune expression in cytotoxic cells, such as NK cells and CD8+ T cells, previous literature has found *CTSW* to be prognostic and correlated with infiltration of both innate and adaptive cell types(*94*). rs61802301, associated with *FCGR2B,* was specifically an cell-type eQTL for B cells; *FCGR2B* is crucial for B cell regulation(*95*). However, variants associated with *TAP2, CTLA-4, FCGR3B, ERAP2,* and DBNDD1 were cell-type eQTLs primarily for myeloid cells, specifically antigen-presenting cells such as macrophages and dendritic cells. *CTLA-4* is expressed on dendritic cells and induces IFNγ production by dendritic cells(*96*). We also observed that targeting *CTSS* resulted in a higher abundance of inflammatory M1 macrophages versus suppressive M2 macrophages, consistent with suppressor myeloid cells impairing effective anti-tumor immunity and ICB response(*97*, *98*). The abundance of innate cell cell-type eQTLs affecting adaptive genes may suggest that a disconnect between innate and adaptive function forms the basis of dysregulated adaptive immunity in the tumor immune microenvironment.

Noteably a SNP-based burden score reproducibly correlated to ICB response across multiple cohorts with melanoma, non-small cell lung cancer (NSCLC) and kidney cancer (RCC). Furthermore, this burden score compared favorably with other popular measures such as tumor mutation burden (TMB) and checkpoint gene expression for predicting binary response category. In renal cell carcinoma the link between tumor mutation burden (TMB) and ICB response is not clear, in contrast to high TMB diseases like melanoma and NSCLC where higher TMB is

associated with better responses(*45*, *99–101*). Possibly, in a setting with low TMB such as in RCC, host genetics have more value as prognostic biomarkers. In the future, germline determinants of the TIME could be integrated into predictors alongside other characteristics of the tumor immune microenvironment that have been found to inform immune response such as TMB, PD-L1 positivity, the number and quality of T cells(*102*), *IFN-γ* response, cytotoxicity scores, T cell activation and T cell exhaustion signatures(*43–45*, *103–110*). Some of these factors require profiling of tumor RNA which is less commonly performed in clinical settings. If germline variants could serve as a proxy for characteristics of the TIME that otherwise require more complex molecular profiling, they could provide an avenue for more cost-effective tools for the clinic.

A potential limitation of the applicability of our approach is that discovery of TIME-SNPs is dependent on the availability of paired genomic and transcriptomic data from tumors, which is currently available only for a few cohorts. Effect sizes associating genetic variants with cellular phenotypes are likely to be larger than those linking genetic variants to diseases(*111–113*), however the number of associations detected may still be limited by available sample sizes and the limited population diversity thereof. We were able to impute a subset of our SNPs into existing immune checkpoint blockade study cohorts that had only exome sequencing, but others falling outside of exonic regions could not be analyzed in this context. Studies focused on tumor exomes and transcriptomes could include genome-wide SNP profiling via arrays or low pass whole genome sequencing to allow more effective integration into future studies of germline genomic variation.

1.1.5 Materials and Methods

1.1.5.1 TCGA Subject Details

The Cancer Genome Atlas (TCGA) consists of tumor and matched normal samples for

over 11,000 patients. The Genomic Data Commons (GDC) legacy archive contains germline data for 11,542 samples from 10,875 unique individuals. Samples with TCGA project IDs: DLBC, LAML, THYM were excluded as they represent cancers derived from immune cells. Pairs of individuals with estimated KING kinship coefficient > 0.177, which represents first-degree relatedness were excluded. TCGA individuals were consented for general research use and no attempts were made to reidentify or contact subjects. Both females and males were included, and sex and individual age were included as covariates. Experimenters were not blinded, and randomization of subjects was not relevant to the study.

1.1.5.2 TCGA Genotypes

Normal (non-tumor) level 2 genotype calls generated from Affymetrix SNP6.0 array intensities using BIRDSUITE (RRID: SCR_001794) software(*114*) were retrieved from TCGA GDC Legacy Portal (accession date: 04/26/2019). In these files, each of 906600 SNPs was annotated with an allele count (0=AA, 1=AB, 2=BB, -1=missing) and confidence score between 0 and 1. Genotypes with a score larger than 0.1 (error rate > 10%) were set to missing and data were reformatted for PLINK (RRID:SCR_001757)(*16*). We discarded 322 SNPs with probe names that did not match the hg19 UCSC Genome Browser (RRID:SCR_005780) Affymetrix track (track: SNP/CNV Arrays, table:snpArrayAffy6). Allele counts were converted to alleles using the definitions in metadata distributed with Affymetrix SNP 6.0 Array Documentation and negative strand genotypes were flipped to the positive strand using PLINK (PLINK, RRID:SCR_001757). Pre-imputation processing of autosomal and X chromosome genotypes consisted of the following steps:

1. SNPs with call rate <90% were removed

2. SNPs with minor allele frequency (MAF) <1% were removed
3. Individuals with genotype coverage <90% were removed
4. Individuals with conflicting gender assignments were flagged
5. Heterozygous haploid SNPs were set to missing.

After applying these filters, the remaining 800644 autosomal and 32809 X chromosome SNPs were input to the secure Michigan Imputation Server(*115*). SNPs were imputed with Minimac3/Minimac4 and European HRC Version r1.1 2016 reference with Eaglev2.3 phasing. Post-imputation processing of genotypes included:

1. SNPs with MAF <1% were removed
2. Autosomal SNPs with Hardy-Weinberg Equilibrium <1e-9 were removed
3. Individuals with high heterozygosity rates (>3 SDs of mean) were removed
4. Pairs of individuals with kinship coefficient > 0.177 (first-degree relatedness) were removed

Rsq values from INFO files were extracted to annotate genotyping quality. The final genotyping data included 8217 individuals and 7,884,718 variants.

1.1.5.3 TCGA Population Stratification

Ancestry filtering was applied using two techniques: (1) k-means clustering and (2) outlier identification. For k-means clustering, TCGA and HapMap Phase III populations were combined. HapMap Phase III genotypes were obtained from the NCBI HapMap ftp site and lifted to hg19 using the liftOver utility(*116*). Genotypes were merged and reduced to a set of 33,675 independent SNPs determined previously(*113*, *116*) through linkage-based filtering using PLINK (RRID:SCR_001757). Pairwise identity-by-state (IBS) between all individuals was calculated and

the resulting IBS matrix was used for PCA analysis. K-means clustering trained on HAPMAP Phase III separated individuals into the following groups: (1) TSI, CEU, (2) JPT, CHD, CHB, (3) MEX, (4) GIH, (5) MKK, (6) YRI, ASW, LWK. This trained model was used to predict groups in TCGA. Cluster (1) were identified as European individuals.

We ran the aberrant R package v1.0 with lambda 20 for outlier identification(*117*). Intersection of k-means clustered individuals and non-outlier individuals from outlier identification analysis was used to include TCGA individuals in the European ancestry discovery cohort.

1.1.5.4 TCGA Phenotype Data

PanCanAtlas RNA data from GDC PanCanAtlas Publications Supplemental Data (https://gdc.cancer.gov/about-data/publications/pancanatlas) was downloaded (access date: 10/14/19). Only primary tumors (barcode: 01A/01B/01C) were considered in our analysis. Corresponding clinical metadata were obtained from the GDC Portal (https://tcga-data.nci.nih.gov/docs/publications/tcga/).

The following phenotypes were extracted or generated from RNA-seq data:

1. **Immunomodulators:** 436 genes used to define immune states from Thorsson et al.
2. **Immune checkpoint molecules:** 78 immune checkpoint stimulatory and inhibitory molecules from Thorsson et al.
3. **Antigen Presentation:** 231 antigen presentation genes from Gene Ontology [GO_REF:0000022]
4. **Immune cell markers:** 60 immune cell type markers from Danaher et al.
5. ***IFN-γ*:** *IFN-γ* genes retrieved from Biocarta [Systematic Name: M18933]

6. ***TGF-β*:** *TGF-β* genes retrieved from Biocarta [Systematic Name: M22085)
7. **Immune states:** Individual level scores for 6 immune states [wound healing, *IFN-γ* dominant, inflammatory, lymphocyte depleted, immunologically quiet, and *TGF-β* dominant] from Thorsson et al.
8. **Immune infiltration levels:** 22 relative immune infiltration estimates from CIBERSORTx(*118*) using LM22 signature matrix.

Phenotypes with greater than 10% zero values were excluded and rank-based inverse normal transformation (**Figure S1)** was applied to each tissue type using below equation (*119*):

qnorm((rank(x,na.last="keep")-0.5)/sum(!is.na(x)))

A total of 733 phenotypes remained for preliminary analyses.

1.1.5.5 Heritability Estimates

Heritability estimates were calculated with the genomic-relatedness-based restricted maximum-likelihood (GREML) approach implemented in GCTA (Genome-wide Complex Trait Analysis)(*120*, *121*). Genetic relationship matrices (GRMs) which measure genetic similarity of unrelated individuals (GRM <0.05) were constructed for the autosomal and X chromosomes for the European cohort. Benjamini-Hochberg false discovery rates (FDR) were calculated using statsmodels(*122*). Immune traits were considered sufficiently heritable if the V(g)/V(p) value was > 0.05 using the full GRM.

As highly polymorphic regions such as HLA and KIR gene regions can inflate heritability estimates, we conducted a 2-state GCTA analysis with separate GRMs for HLA/KIR regions (HLA chr6:28,477,797-33,448,354, KIR chr19:55,228,188-55,383,188) and with the rest of the genome excluding HLA/KIR regions. Age and sex were included as covariates. An FDR < 0.05

was used to identify SNP-heritable IP components from 2-state analysis. If an IP component had high SNP heritability using the HLA/KIR GRM, a conditional GWAS analysis was conducted; otherwise, a standard GWAS analysis with Bonferroni-corrected suggestive p-value threshold was conducted. Ultimately, 140 IP components outside of the HLA/KIR regions and 17 IP components within the HLA/KIR regions were identified.

1.1.5.6 Principal Component Analysis

157 SNP-heritable components were analyzed using sklearn. IP component values were scaled by Sklearn Standard Scaler and used for principal component analysis (PCA). Ordinary least squares (OLS) regression was performed with 157 IP components and principal components, wherein the beta coefficient represents the degree of change in principal component for every unit change in IP component. P-values indicate whether a coefficient was significantly different from 0.

1.1.5.7 GWAS Analysis

The GLM method in PLINK (RRID:SCR_001757) was used to conduct association analyses with IP components. All associations were adjusted for covariates of age, sex and the first ten principal components. Gene expression values, CIBERSORTx relative infiltration estimates, and immune state scores were inverse-rank normalized by tissue type to control for tissue-type expression effects. Significant associations were identified with the PLINK (RRID:SCR_001757) clumping method using the primary suggestive threshold corrected for the number of phenotypes tested(*123*) (1x10-5/140) using a kb threshold of 500, and an R^2 threshold of 0.5.

To determine if variants had been implicated in previous cancer GWAS studies, variants

were input into the LDlink server (https://ldlink.nci.nih.gov/?tab=ldtrait) using parameters "EUR" population, an R^2 threshold of 0.5 and base pair window of 500kb(*36*, *37*). We also retrieved the Vanderbilt PheWAS catalog(*38*) and any TIME-SNPs in high linkage disequilibrium ($R^2 > 0.5$) with Vanderbilt PheWAS catalog cancer risk SNPs were included as cancer risk variants. Lastly, we assessed SNPs by PheWAS analysis in the UK Biobank (detailed below).

1.1.5.8 Conditional HLA analysis

The PLINK (RRID:SCR_001757) GLM method was used to run stepwise conditional analysis for identification of independent HLA associations(*20*). The most significant initial associations detected with HLA region phenotypes by standard GWAS analysis were incorporated as covariates in the subsequent round. Specifically, we re-ran the analysis with chromosome 6 variants including the most significant SNP (lowest p-value in the previous round) as a covariate. Analysis was conducted until no SNPs with Bonferroni-corrected p-value < ($1x10^{-5}/17$) remained.

1.1.5.9 HLA Allele Specific Expression

TCGA tumor specific RNA BAM files were downloaded from the GDC on 07/16/2019. The HLApers(*124*) kallisto-based pipeline was used with gencode.v30 annotations(*125*). Default parameters were used and the two alleles with the highest calculated expression were retained for each HLA gene if there were more than 2 alleles reported. The top 2 highest expressed HLA alleles for each gene were averaged for input into SNP analyses. If expression for at least two alleles was not calculated, expression was set as missing for the sample. Only primary samples (01A/01B/01B) were considered for analysis. Summed HLA allele specific expression was inverse-rank normalized by cancer type and used for downstream analyses.

1.1.5.10 Literature TIME-SNPs

We compiled existing germline variants associated with the tumor immune microenvironment (TIME) or ICB response from the literature. We collected 14 studies with their descriptions below:

1. Kogan et al. *JCI* 2018: Discovery of *FGFR4* germline variant which enhances *STAT3* activity impeding CD8 T cell infiltration
2. Queirolo et al. *Front Immunol* 2017: Investigation of 6 *CTLA-4* SNVs in 173 metastatic melanoma patients with overall response and survival information
3. Uccellini et al. *J Transl Med* 2012: *IRF5* polymorphism was associated with non-response to adoptive therapy with TILs
4. Bedognetti et al. *Br J Cancer* 2013: *CXCR3* and *CCR5* genetic polymorphisms were evaluated for expression of respective ligands and TIL migration
5. Lim et al. *PNAS* 2018. Systematic identification of germline genetic polymorphisms associated xCell cell type gene signatures (gsQTLs) in TCGA.
6. Shahamatdar et al. *Cell Reports* 2020. Systematic identification of germline genetic polymorphisms associated with immune infiltration in TCGA.
7. Ostendorf et al. *Nat Medicine* 2020. Identification of *APOE2* and *APOE4* germline variants associated with melanoma progression and ICB response in mice.
8. Zhang et al. *Front Genet* 2019. Identification of breast-cancer associated variant modulating *CTSW* expression
9. Sayaman et al. *Immunity* 2020. Systematic identification of germline variants associated with 33 immune traits including leukocyte subsets, adaptive receptor, immune expression

signatures.

10. Yoshida et al. *Eur J Cancer* 2021. Identification of 2 *PD-L1* variants associated with survival outcomes in advanced non-small-cell lung cancer patients.
11. Kula et al. *Exp Mol Pathol* 2020. Review of 10 *PD-L1* genetics variants.
12. Salmaninejad et al. *Immunogenetics* 2018. Review of 5 frequently studied *PD-1* genetic variants.
13. Sasaki et al. *Mol Clin Oncol* 2014. Characterization of *PD-1* promoter variant and association with survival in non-small cell lung cancer.
14. Tang et al. *Int J Clin Exp Med* 2015. Characterization of 3 *PD-1* variants and association with cancer risk.

For Sayaman et al., 598 significant associations were identified, 520 of which were within the MHC II region. To identify independent Sayaman et al SNPs, we performed linkage disequilibrium-based clumping with the same parameters used for our analysis. After clumping, 55 independent Sayaman et al SNPs remained.

1.1.5.11 GREGOR

GREGOR was used to analyze SNP enrichment at epigenetic features. We obtained 479 bed files for 11 histone experiments and 52 cell types from ENCODE (downloaded on May 3, 2020**).** Only "stable peaks" and "replicated peaks" files were kept for analysis. If more than 1 bed file for a cell type and histone mark were available, the files were combined.

In addition, 323 bed files for 12 transcription factor binding experiments and 12 cell types were downloaded from ENCODE (RRID:SCR_015482**)** on August 4, 2020. Only "optimal IDR thresholded peaks" and "conservative IDR thresholded peaks" files were kept. If more than 1 bed

file for a cell type and transcription factor were available, the files were combined.

GREGOR (RRID: SCR_009165) was run with EUR Reference files made from the 1000 Genomes Project data with an LD window size of 1MB and LD $R^2 > 0.7$. Enrichment ratios were calculated by taking the difference between observed and expected number of SNPs and dividing by the expected number of SNPs. Any files with Audit errors were excluded.

1.1.5.12 Variant Annotation

Variants were annotated with VEP (Variant Effect Predictor)(*126*) with default parameters and the GRCh37 reference genome. Coding variants were mapped to protein sequences using the Uniprot GFF file.

1.1.5.13 Cell-Type eQTL Analysis

We followed the GTEx approach for cell type interaction eQTL discovery(*17*). We ran a linear regression model with an interaction term accounting for interactions between genotype and cell type enrichment from xCell(*127*):

$$p \sim g + i + g \circ i + C$$

where p is the IP component vector, g is the genotype vector, i is the inverse normal transformed by tissue type xCell enrichment score(*127*), and the interaction term $g \circ i$ corresponds to pointwise multiplication of genotypes and cell type enrichment scores. The same covariates, denoted by C, were used as in the regular immune microenvironment GWAS analysis. Benjamini-Hochberg FDR was calculated for the beta coefficient of the interaction term and variants with FDR < 0.1 were identified as significant.

DICE expression quantitative trait loci (eQTLs) were obtained at https://dice-

database.org/. Methods associated with DICE eQTL discovery are published in Schmiedel et al(*128*).

1.1.5.14 UK Biobank

UK Biobank subjects were subsetted into separate ethnic-racial groups following continental ancestry prior to analysis. The sub-setting was performed to generate homogenous groups and reduce potential admixture bias in the genetic analyses. To identify the European-ancestry samples, we started with directly called genotype data and identified a set of overlapping SNPs with 1000 Genomes Project and AWS (RRID:SCR_008801) (1KG) population and then merged them together. Next, we pruned the SNP set so remaining SNPs were in linkage equilibrium using PLINK (PLINK, RRID: SCR_001757)(*16*). flashpca was used to calculate principal components for 1KG SNPs(*129*). The UK Biobank samples were projected onto 1KG space using flashpca. To identify subjects of European ancestry, we utilized Aberrant to generate clusters with a broad set of lambda values (clustering thresholds) and checked that the cluster included all 1KG subjects of European ancestry and maximized the total number of UK Biobank subjects (lambda=8.2)(*117*). Finally, we compared the self-reported race/ethnicity of subjects within this cluster and removed samples that were discordant. We identified 454,487 subjects of European ancestry. To identify the unrelated samples from the finalized European list, we used the relatedness file provided by UK Biobank and a custom script was used to select unrelated samples while maximizing sample counts. The final European unrelated set included 382,841 subjects.

Variant dosages extracted from imputed UK Biobank BGEN files were used for PheWAS analysis with PLATO v2.0.0(*130*). ICD10 diagnosis codes associated with neoplasms and immune disorders were collapsed according to level-1 groupings used by UK Biobank resulting in a total

of 24 groups. For example, C00-C14 is one of the groups containing ICD10 codes associated with malignant neoplasm of lip, oral cavity, and pharynx. Individuals with diagnosis code in a group were coded as 1, with the remaining individuals coded as 0. Logistic regression was conducted with UK Biobank binary files containing HLA-immune variants, logistic phenotype file, and age, sex, and principal components 1-10 as covariates. Pvalues were Benjamini-Hochberg FDR adjusted.

1.1.5.15 Survival Analysis

Kaplan-Meier analysis of immune microenvironment associations were conducted with overall and progression-free survival retrieved from Liu et al.(*131*) by cancer type using the lifelines package. As recommended by Liu *et al,* TCGA cancer types, TGCT and PCPG, were excluded as survival data did not meet quality standards. TCGA individuals were divided into 3 groups based on genotype calls: minor allele homozygotes, heterozygotes and major allele homozygotes. Significance was determined using the logrank test between minor allele and major allele homozygotes. Only SNPs with at least 1% minor allele frequency in each cancer type and more than 1 minor allele homozygous individual were considered for analysis. Logrank p values were corrected using the Benjamini-Hochberg method from the statsmodel package and only variants with FDR < 0.1 were considered.

1.1.5.16 Immune Checkpoint Blockade (ICB) Cohort Genotypes

Raw fastq files were obtained for the following immune checkpoint trials: Hugo et al. 2016 (SRA accession: SRP090294, SRP067938; Cancer: melanoma)(*41*), Van Allen et al. 2015 (SRA accession: SRP011540, Cancer: melanoma)(*42*), Miao et al. 2018 (SRA accession: SRP128156,

Cancer: clear cell renal carcinoma)(*132*), Riaz et al. 2017 (SRA accession: SRP095809, SRP094781; Cancer: melanoma)(*44*), Rizvi et al. 2015 (SRA accession: SRP064805, Cancer: non-small cell lung cancer)(*43*), Snyder et al. 2014 (SRA accession: SRP072934, Cancer: melanoma)(*45*). Reads were aligned to UCSC hg19 coordinates using BWA (RRID:SCR_010910) v0.7.17-r1188(*133*). Reads were sorted by SAMTOOLS (RRID:SCR_002105) v0.1.19(*134*, *135*), marked for duplicates with Picard Tools (RRID:SCR_006525) v2.12.3 and recalibrated with GATK (RRID:SCR_001876) v3.8-1-0(*136–138*). Germline variants were called from sorted BAM files using DeepVariant v0.10.0-gpu(*139*, *140*). The final immunotherapy cohort consisted of 68 clear cell renal carcinoma, 276 melanoma and 34 non-small cell lung cancer patients.

To evaluate the quality of SNP imputation from whole exome data, we took advantage of the TCGA having both. Of the 1,322,586 variants available from DeepVariant analysis of immunotherapy cohort, 225,000 were available in TCGA imputed data. We extracted these 225,000 variants from TCGA and input into the Michigan Imputation Server (reference panel: HRC, phasing: Eagle). We compared genotypes from whole-exome calls vs. original Affymetrix-based TIME-SNP calls. Variants with >5% mismatches in genotype calls, minor allele frequency < 5% in any cohort or imputation accuracy ($R^2 < 0.3$) were excluded. Only variants with at least 5% frequency in all 4 melanoma cohorts used for discovery analysis were considered for ICB analysis, leaving 525 SNPs.

Population stratification analysis was conducted by taking overlapping variants between TCGA and ICB cohorts. Variants with MAF differences > 0.1% were excluded resulting in 3612 frequency-concordant variants. PLINK IBD analysis was conducted, and top 10 principal components were included in association analysis.

1.1.5.17 Immune Checkpoint Blockade (ICB) Response Analysis

Subject phenotypes were downloaded from supplementary information of ICB trial publications. Four melanoma cohorts were used as the discovery cohort for ICB-associated variants, while Miao et al renal cell carcinoma and Rizvi et al non-small cell lung cancer cohorts were used for validation. Response phenotypes were determined from iRECIST criteria(*141*). Patients were categorized as responders if they had iRECIST criteria: CR (complete response), PR (partial response) and SD (stable disease). Genome-wide association studies (GWASs) were conducted for ICB responders within each ICB-cohort using PLINK (PLINK, RRID:SCR_001757). Age, sex and the top 10 principal components were included in the logistic analysis as covariates. We then used METAL(*48*) with a sample size weighting scheme to perform a pan-study melanoma meta-analysis for ICB response.

Eleven variants (FDR < 0.5) were selected for construction of a germline burden score. Variant genotypes were aligned such that the allele associated with higher odds of responder status were used for construction of the burden score, which was a simple count of the number of ICB-associated variants for each individual.

1.1.5.18 Immune Checkpoint Blockade (ICB) Response RNA-seq

FASTQ/BAM files were downloaded for 33 RCC and 120 melanoma patients. BAM files were converted to FASTQ using bam2fq(*135*). Unpaired reads were removed using fastq pair(*142*). Paired reads were aligned with STAR (RRID:SCR_004463) v2.4.1d(*143*) to GRCh37 reference alignment. RSEM(*144*) was used for transcript quantification. TPM values were log2 transformed for analyses. Differential gene expression analysis between responders and non-responders from cohorts Riaz et al. 2017, Hugo et al. 2016, Miao et al. 2018, and Van Allen et al.

2015 was performed using the DESeq2[143] package in R. Cohort was included as a covariate when calculating top differentially expressed genes.

1.1.5.19 Mouse experiments

Wild-type C57BL/6 (RRID:IMSR_JAX:000664) were purchased from The Jackson Laboratory. All the animal studies were approved by the Institutional Animal Care and Use Committee (IACUC) of university of California, San Diego, with protocol ASP #S15195. Mice at Moores Cancer Center, UCSD are housed in micro-isolator and individually ventilated cages supplied with acidified water and fed 5053 Irradiated Picolab Rodent Diet 20 lab diet. Temperature for laboratory mice in our facility is mandated to be between 65–75 ° F (~18–23 °C) with 40–60% humidity. All animal manipulation activities are conducted in laminar flow hoods. All personnel are required to wear scrubs and/or lab coat, mask, hair net, dedicated shoes, and disposable gloves upon entering the animal rooms. $2x10^5$ MC38 (RRID:CVCL_B288) cells were transplanted into the flank of 8-10 female C57Bl/6 (RRID:IMSR_JAX:000664) mice, aged 7-8 weeks. Where indicated, when tumors reached 100 mm^3, mice were randomized and treated with anti-PD-1 (10mg/kg i.p., Bio X Cell Cat# BE0146, RRID:AB_10949053, clone RMP1-14), CTSS inhibitor (5mg/kg, i.p., APEx Bio) or isotype control antibody. Treatments were given 3 times a week. Where indicated (or when control-treated mice succumbed to tumor burdens, as determined by the ASP guidelines), mice were euthanized and tumors were taken for flow cytometric analysis. MC38 cells were not screened using STR profiled on site.

1.1.5.20 Reagents

PD-1 antibody (clone RMP1-14, Bio X Cell Cat# BE0146, RRID:AB_10949053) and

isotype antibody (catalog #BE0091) were purchased from Bio X Cell.

1.1.5.21 RT-PCR

RNA from MC38 (RRID:CVCL_B288) tumors was extracted using the RNeasy Mini Kit (Qiagen catalog #74104). 500ng of RNA per reaction was used to prepare cDNA with the SuperScript™ VILO™ cDNA Synthesis Kit (ThermoFisher Scientific) following manufacturer's instructions. The cDNA was used to set up the RT-PCR reaction with 4 technical replicates per tumor with the Fast SYBR™ Green Master Mix (ThermoFisher Scientific) according to manufacturer's instructions. PCR quantification was conducted using the $2^{-\Delta\Delta CT}$ method and normalized to the housekeeping gene β-actin.

RNA-seq and CIBERSORTx infiltration estimates for M4 melanoma mouse model were obtained from GEO accession (GSE144946). Responders were mice whose size at harvest was smaller than the last dose of anti-CTLA-4. RNA-seq counts were converted to TPM and log2 normalized.

1.1.6 Figures

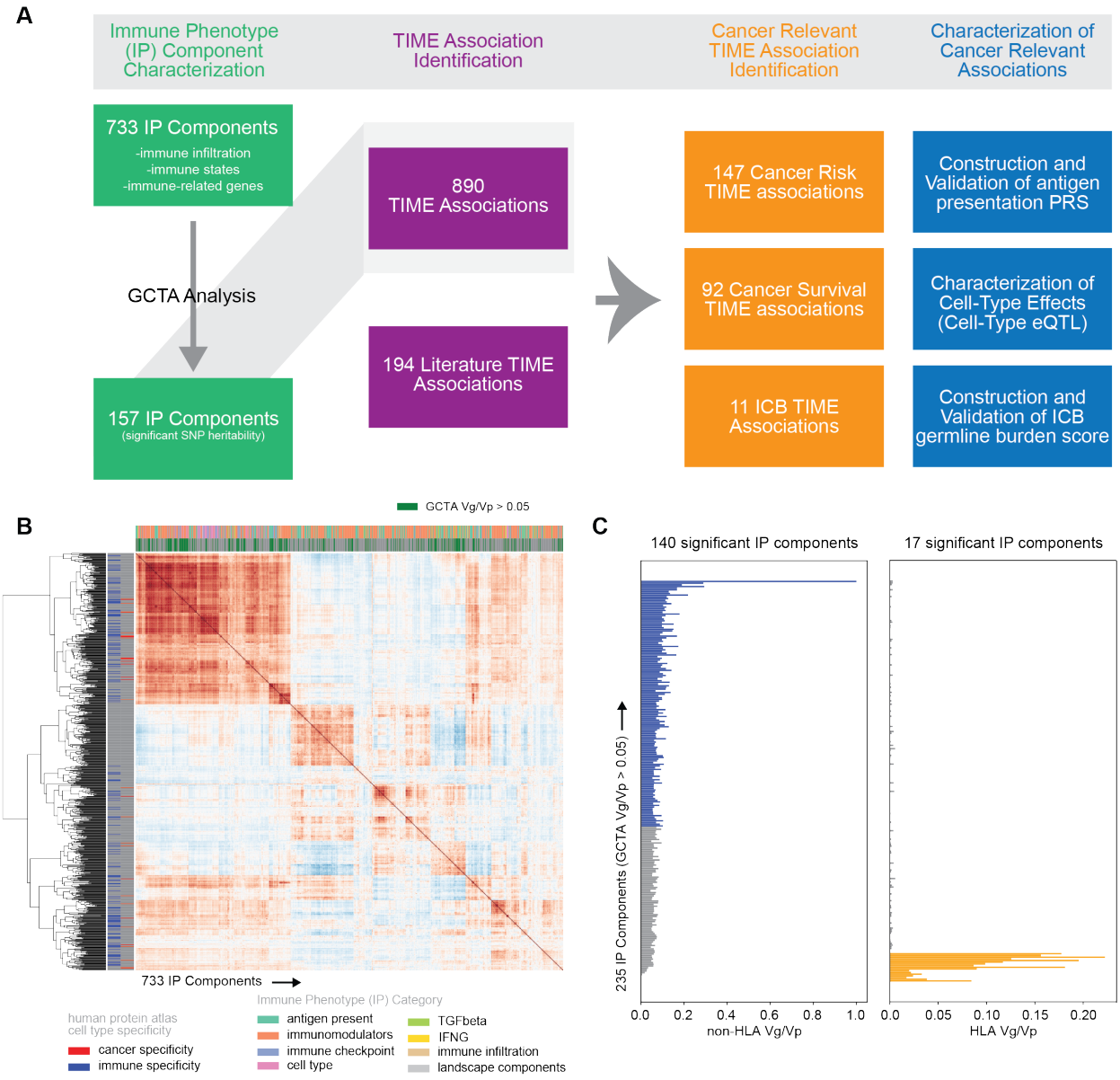

Figure 1.1 Characterization of Tumor Immune Microenvironment (TIME) (A) Overview of the TIME germline analysis. **(B)** Clustermap depicting 733 IP components and their pairwise correlation across 30 tumors in the TCGA. **(C)** Horizontal barplot of variance in phenotype explained by variance in genotype (Vg/Vp) for 235 IP components estimated separately genome-wide excluding the HLA locus (left panel) and using only the HLA locus (right panel).

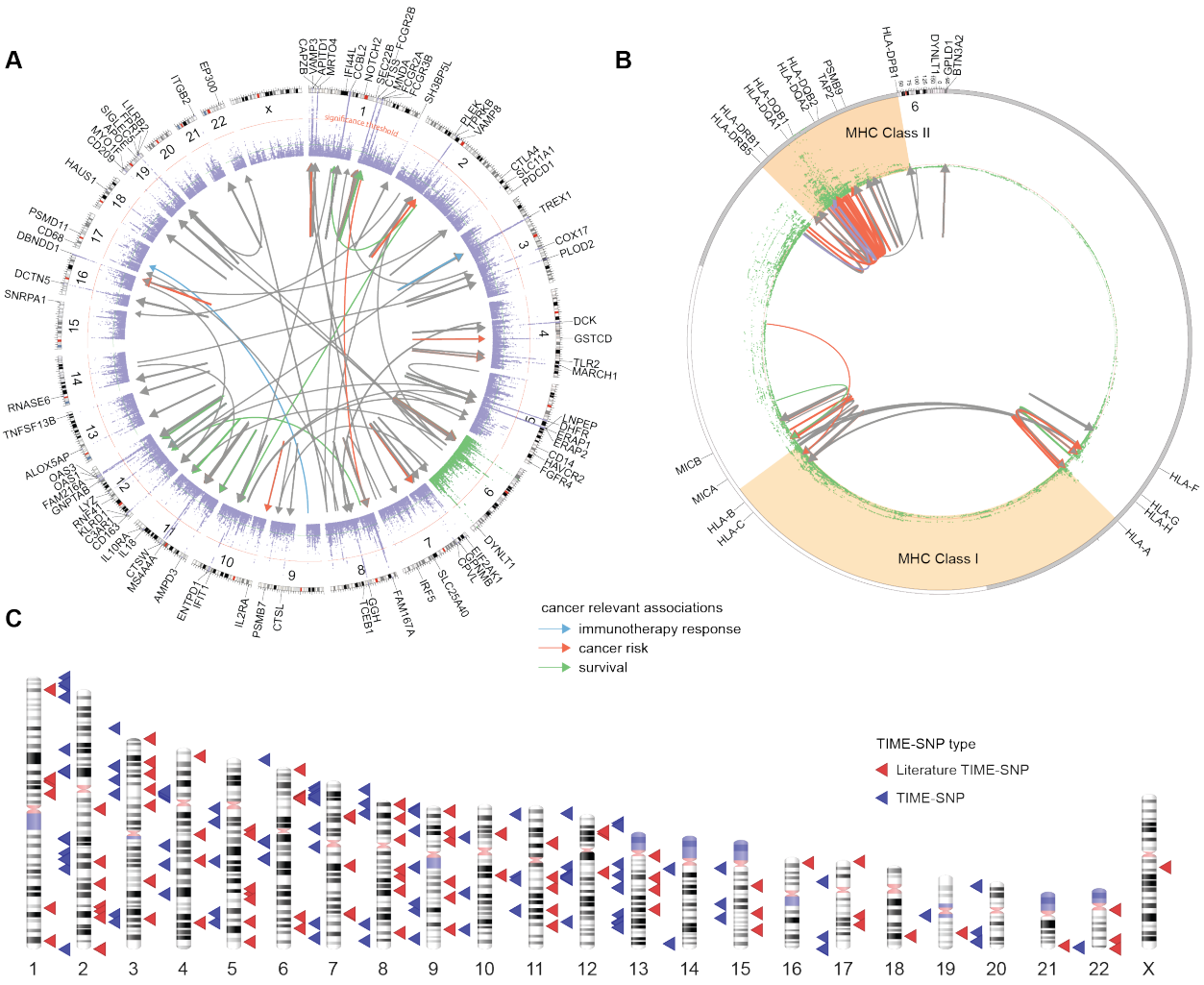

Figure 1.2 Detecting Putative Germline Modifiers of the Tumor Immune Microenvironment **(A)** Locuszoom plot summarizing 890 associations between germline SNPs and 93 IP components. Outer ring represents locations of all 157 tested IP components. Links are colored if implicated in cancer risk (orange), survival (green) or immunotherapy response (blue) **(B)** Significant associations between SNPs and 17 IP components in the HLA region detected through conditional GWAS analysis for effects on gene expression using either a basic alignment to the reference genome (conditional), or allele specific expression obtained by aligning to a patient-specific HLA reference allele set. **(C)** Ideogram plot of TIME-SNPs implicated by our discovery analysis (red) and literature curation (blue).

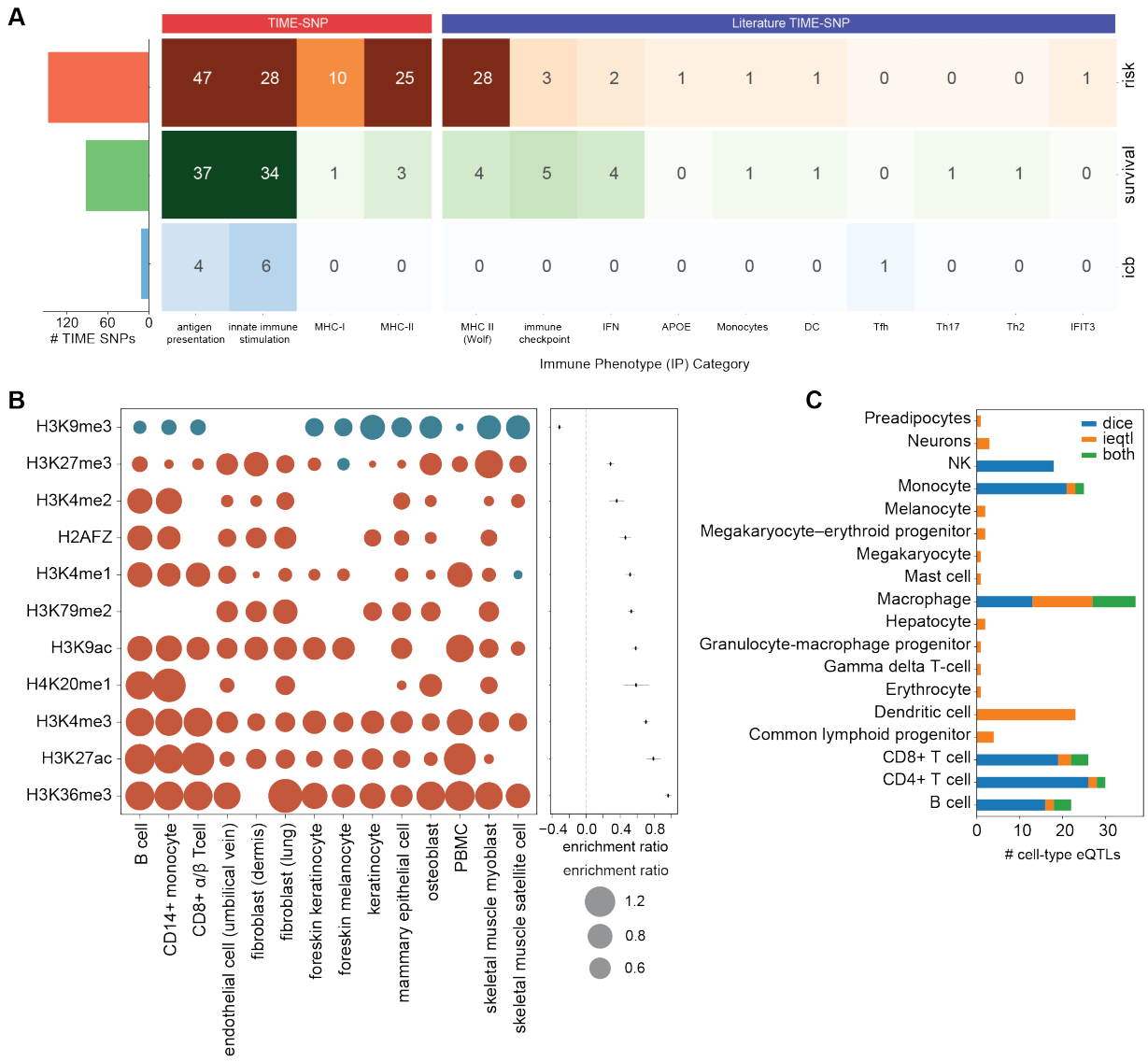

Figure 1.3 Identification and Characterization of TIME-SNPs Related to Cancer Outcomes **(A)** Clustermap of cancer relevant associations with barplot of TIME-snps implicated in cancer risk, survival and immunotherapy response. **(B)** Mean enrichment ratio of immune microenvironment variants in histone marks with corresponding enrichment ratios in specific cell types. **(C)** Barplot of cell-type specific TIME-SNPs implicated by DICE and ieQTL analysis.

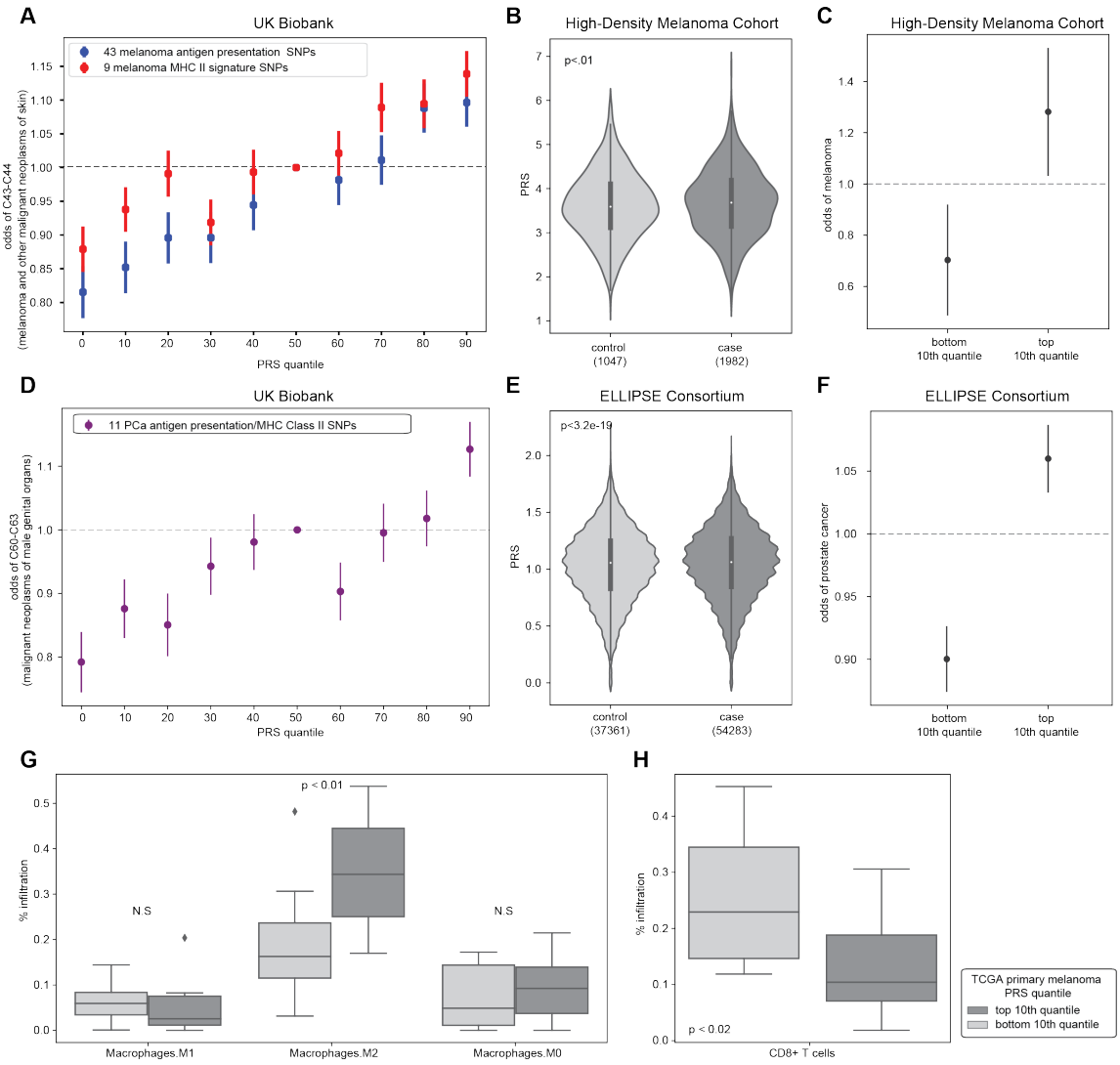

Figure 1.4 TIME-SNPs Underlying Antigen Presentation Stratify Melanoma and Prostate Cancer Risk (A) UK Biobank quantile plot of polygenic risk score (PRS) constructed from antigen presentation gene TIME-SNPs and Sayaman et al. MHC II signature TIME-SNPs associated with melanoma risk. **(B)** Violinplot of PRS constructed from antigen presentation genes and MHC II signature TIME-SNPs in melanoma cases and controls in High Density melanoma cohort. **(C)** Odds of melanoma risk among individuals in the top and bottom 10th quantile of PRS in ELLIPSE consortium. **(D)** UK Biobank quantile plot of PRS constructed from antigen presentation gene TIME-SNPs and Sayaman et al. MHC II signature TIME-SNPs associated with prostate cancer risk. **(E)** Violinplot of PRS constructed from antigen presentation genes and MHC II signature TIME-SNPs in prostate cancer cases and controls in ELLIPSE Consortium. **(F)** Odds of prostate cancer risk among individuals in the top and bottom 10th quantile of PRS in ELLIPSE consortium **(G)** Boxplot of macrophage infiltration in primary TCGA SKCM (melanoma) in top and bottom 10th quantile of melanoma antigen presentation TIME-SNP PRS. **(H)** Boxplot of CD8+ T cell infiltration in TCGA SKCM (melanoma) in top and bottom 10th quantile of melanoma antigen presentation TIME-SNP PRS.

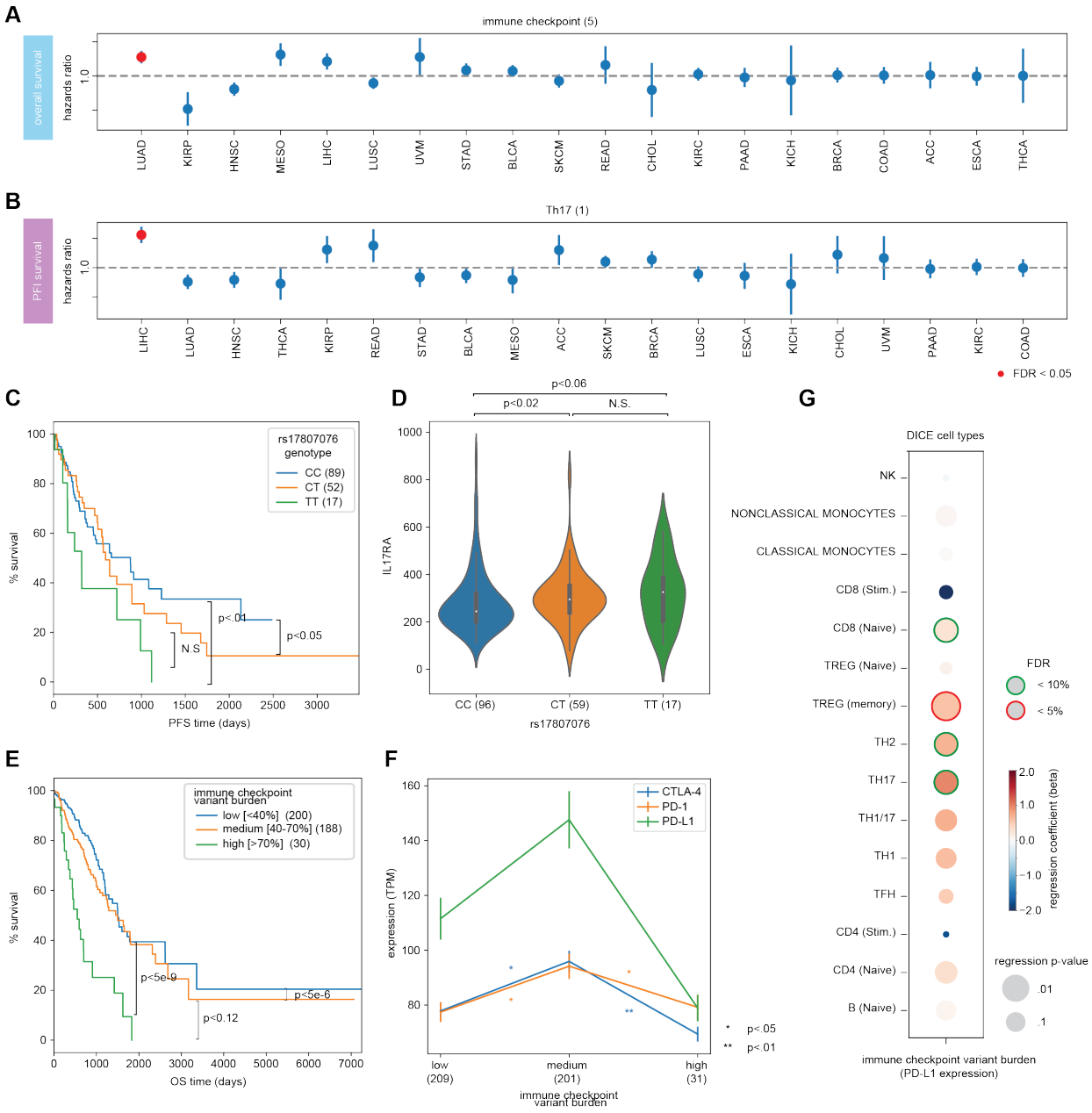

Figure 1.5 Variants Underlying Immune Evasion are Associated with Cancer Survival (A) Cox Proportional-Hazards odds ratios for immune checkpoint burden score (constructed from 5 immune checkpoint variants) with overall survival separated by TCGA cancer type. **(B)** Cox Proportional-Hazards odds ratios for rs17807076 and progression-free survival separated by TCGA cancer type. **(C)** Progression-free survival Kaplan-Meier curve based on rs17807076 genotype in TCGA LIHC. **(D)** Violin plot of rs17807076 genotype and bulk RNA-seq expression of *IL17RA* in TCGA LIHC. **(E)** Overall survival Kaplan-Meier curve based on burden of 5 immune checkpoint variants (immune checkpoint variant burden) in TCGA LUAD. **(F)** Average expression of immune checkpoint molecules, *PD-L1, PD-1* and *CTLA-4*, stratified by immune checkpoint variant burden **(G)** Association of *PD-L1* expression with immune checkpoint variant burden in 15 immune cell types from DICE.

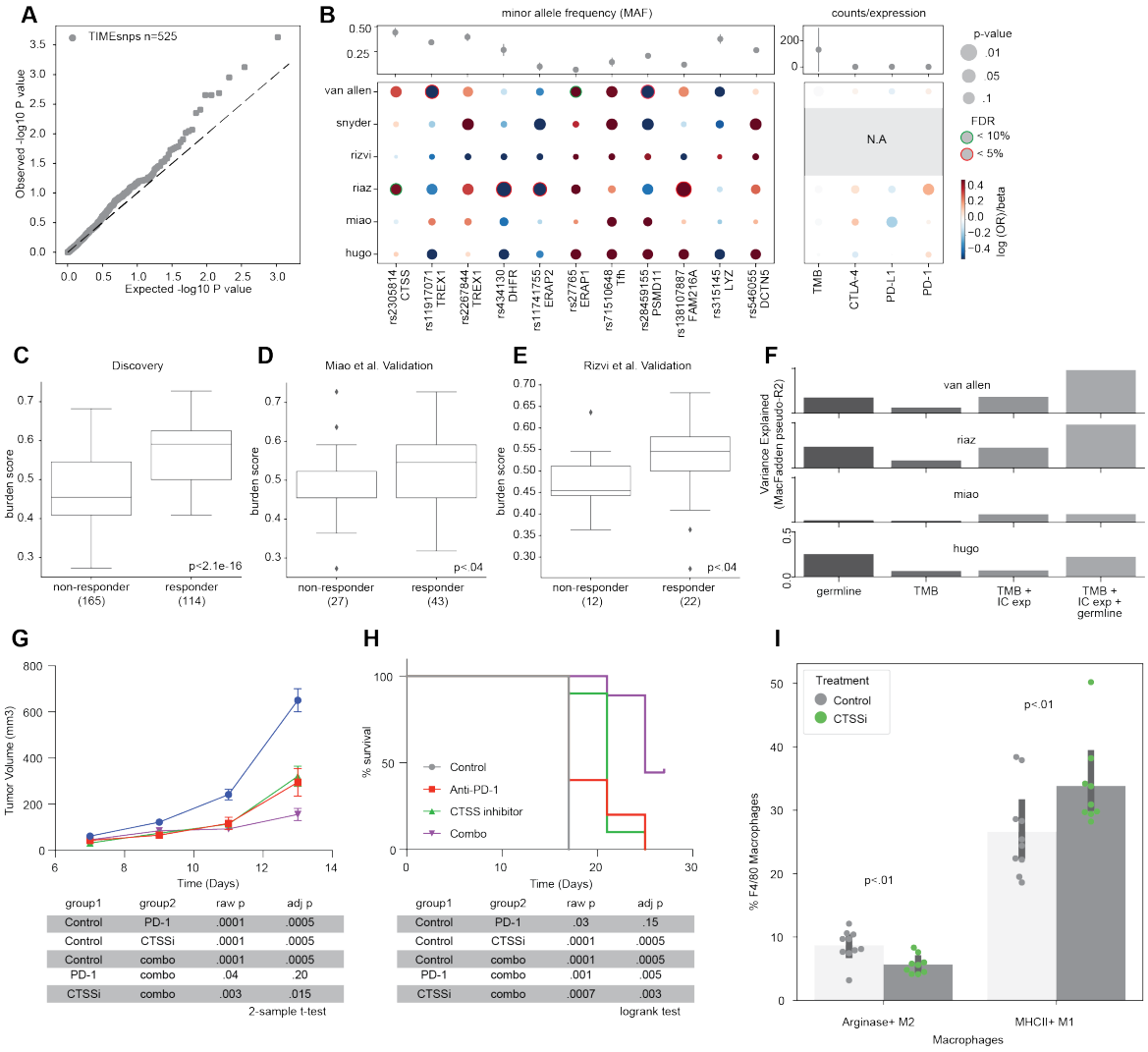

Figure 1.6 TIME-SNPs Implicate Targets for Modulating Immune Responses (A) Q-Q plot of METAL meta-analysis sample-size weighted association with immunotherapy response in Riaz et al., Snyder et al., Hugo et al. and Van Allen et al. melanoma discovery cohort. **(B)** Grid plot of log odds ratio of variants with responder status in 6 ICB cohorts with beta coefficients of classic ICB biomarkers (TMB, *PD-L1, PD-1, CTLA-4*) association with responder status. **(C)** Boxplot of burden score constructed from 11 significant ICB variants (FDR < 0.5) in discovery melanoma cohort. **(D)** Boxplot of burden score constructed from 11 significant ICB variants (FDR < 0.5) in Miao et al validation cohort. **(E).** Boxplot of burden score constructed from 11 significant ICB variants (FDR < 0.5) in Rizvi et al validation cohort. **(F)** Pseudo-R^2 (variance explained) with germline burden score only, TMB only, expression of *PD-L1, PD-1* and *CTLA-4*, and germline burden score with TMB and expression of *PD-L1, PD-1* and *CTLA-4*. **(G)** Tumor growth curve for C57BL/6 mice implanted with MC38 treated with anti-PD-1, anti-CTSS, and combination of anti-PD-1 and anti-CTSS. **(H)** Survival curve for C57BL/6 mice implanted with MC38 treated with anti-PD-1, anti-CTSS, and combination of anti-PD-1 and anti-CTSS. **(I)** Barplot of the proportion of F4/80 Macrophages that are Arginase$^+$ M2 macrophages and MHCII$^+$ M1 macrophages respectively for MC38 tumors treated with anti-CTSS compared to controls

1.1.7 Tables

Table 1.1 TCGA Cancer Types

TCGA Cancer type abbreviation	Cancer
LUSC	Lung squamous cell carcinoma
LUAD	Lung adenocarcinoma
KIRC	Kidney renal cell carcinoma
KIRP	Kidney renal papillary cell carcinoma
COAD	Colon adenocarcinoma
BRCA	Breast invasive carcinoma
READ	Rectum adenocarcinoma
UCEC	Uterine corpus endometrial carcinoma
LIHC	Liver hepatocellular carcinoma
THCA	Thyroid carcinoma
BLCA	Bladder urothelial carcinoma
STAD	Stomach adenocarcinoma
PRAD	Prostate adenocarcinoma
HNSC	Head and Neck squamous cell carcinoma
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma
SARC	Sarcoma
SKCM	Skin cutaneous melanoma
PAAD	Pancreatic adenocarcinoma
ESCA	Esophageal carcinoma
KICH	Kidney chromophobe
PCPG	Pheochromocytoma and Paraganglioma
CHOL	Cholangiocarcinoma

1.1.8 Supplementary Figures

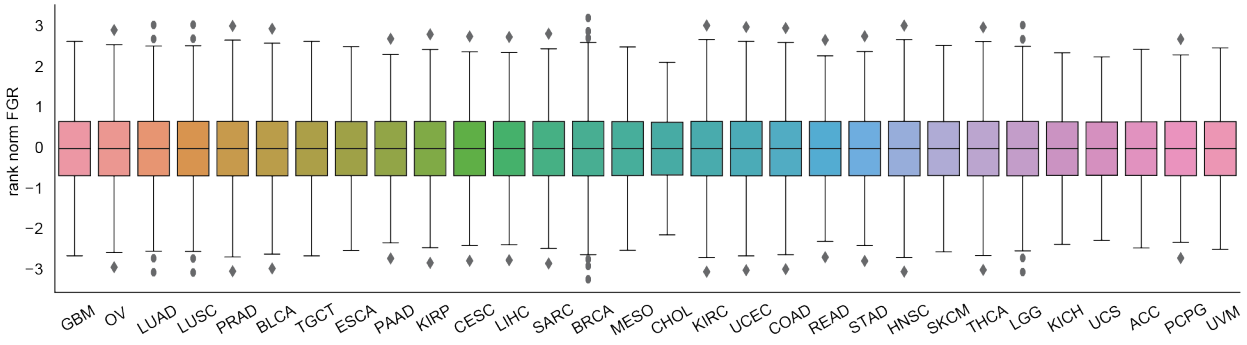

Figure S1.1 Characterization of Tumor Immune Microenvironment (A) Boxplot of FGR expression values after inverse-rank normalization by cancer type. **(B)** Adjacency matrices created from Pearson Correlation matrices amongst immune phenotypes were used to construct highly related immune phenotypes hubs. Phenotypes from different literature sources are colored differently. **(C)** Boxplot of CIBERSORTx infiltration across TCGA individuals. **(D)** Barplot of V(g)/V(p) values for 137 heritable IP components immune phenotypes using a two-state GCTA model: HLA/KIR polymorphic regions (blue), rest of the genome (orange).

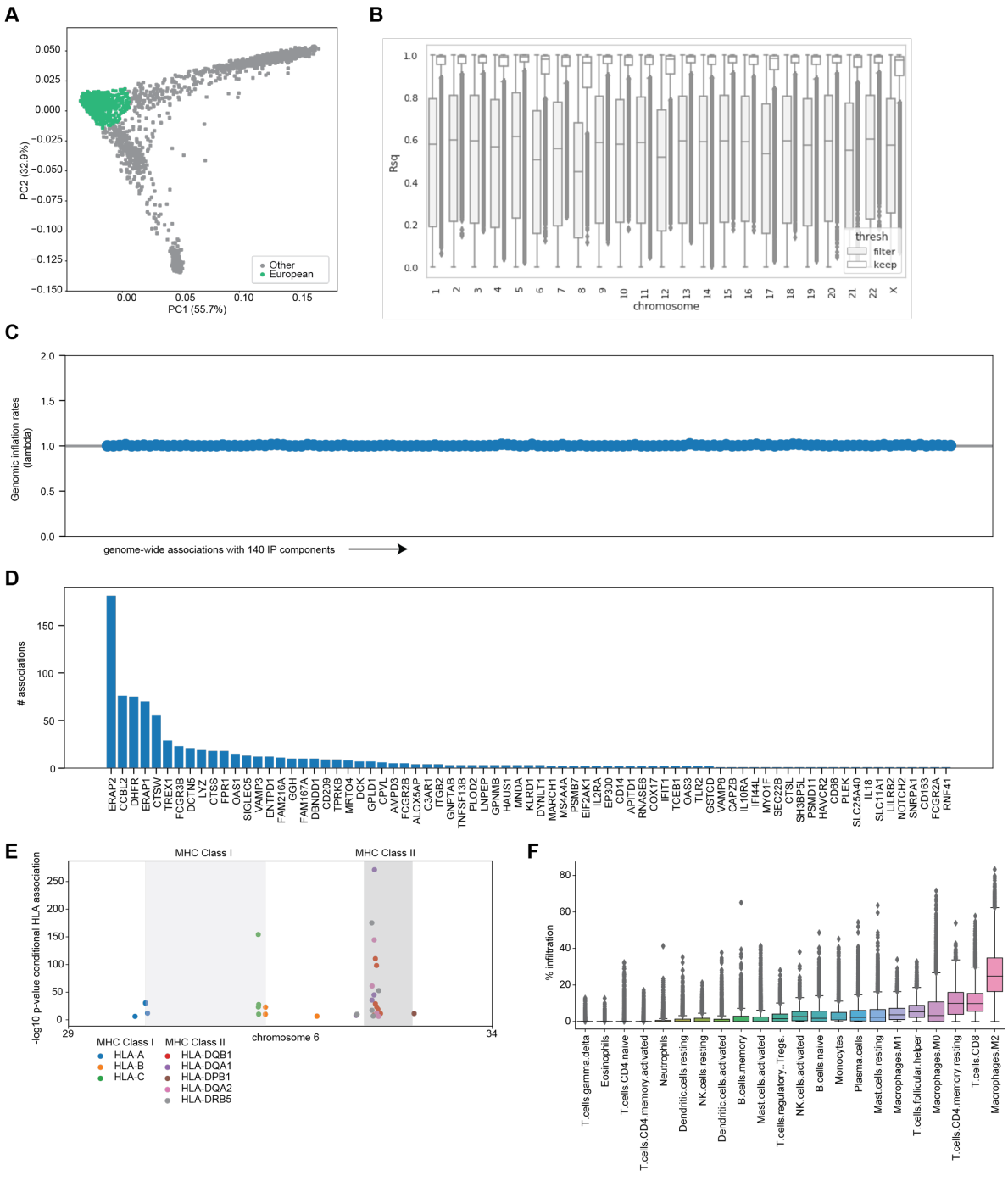

Figure S1.2 Detecting Putative Germline Modifiers of the Tumor Immune Microenvironment (A) PCA analysis of TCGA genotypes; green indicates European discovery cohort. **(B)** Boxplot of Rsq distribution of imputed genotypes across chromosomes. **(C)** Plot of inflation factors (lambda) for associations using PancanAtlas data. **(D)** Barplot of number of associations for each immune phenotype (IP) components. **(E)** Scatterplot of HLA associations based on chromosome 6 genomic location. **(F)** Boxplot of CIBERSORTx infiltration across TCGA individuals.

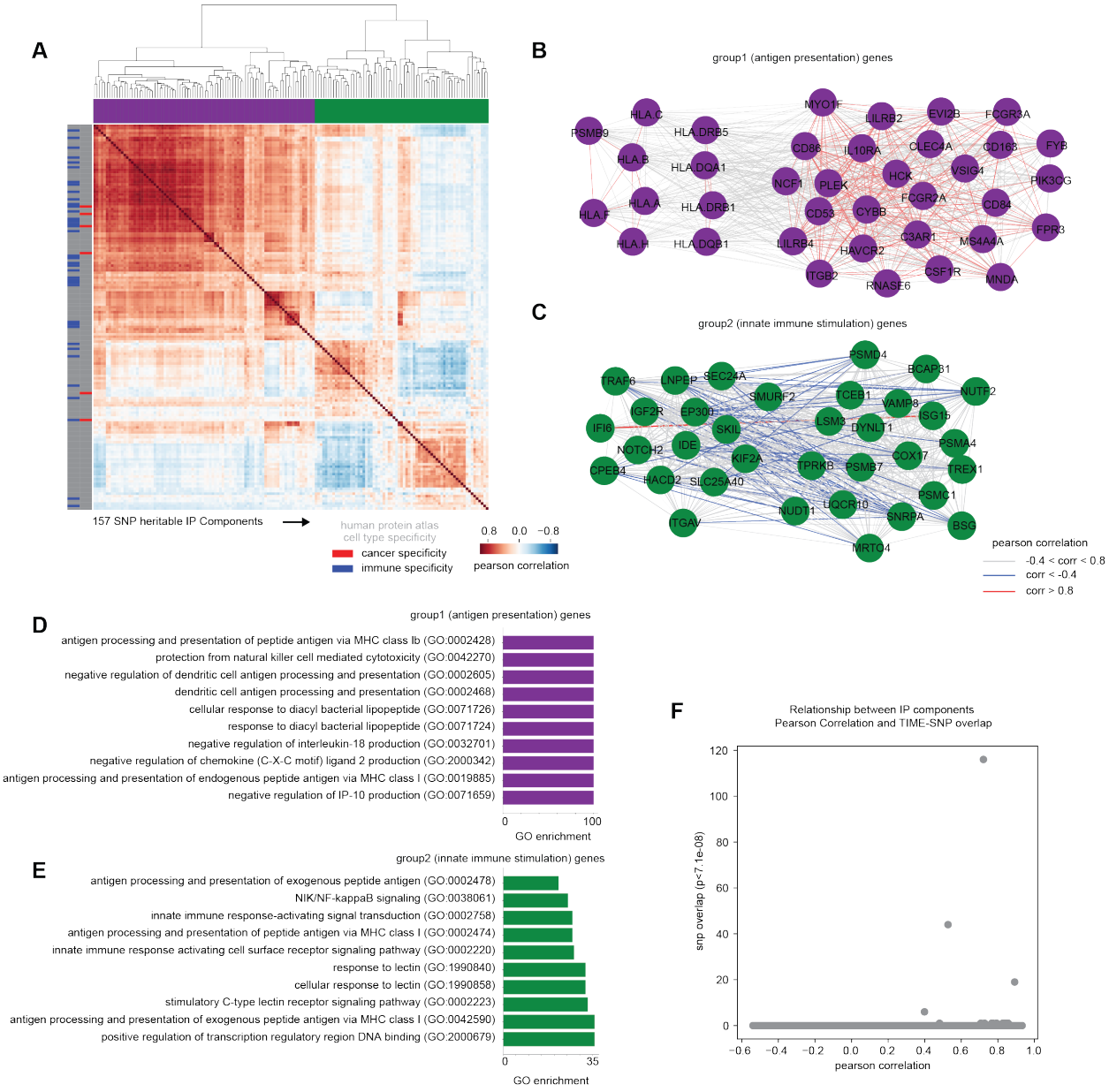

Figure S1.3 Detecting Putative Germline Modifiers of the Tumor Immune Microenvironment (A) Clustermap depicting 157 SNP-heritable IP components from two-state GCTA analysis and their pairwise correlation across 30 tumors in the TCGA. **(B)** Network plot of group 1 (antigen presentation) genes from pearson correlation Clustermap analysis. **(C)** Network plot of group 2 (innate immune stimulation) genes from pearson correlation Clustermap analysis **(D)** Top 10 GO enrichment terms and enrichment values from group 1 (antigen presentation) genes. **(E)** Top 10 GO enrichment categories and enrichment values from group 2 (innate immune stimulation) genes. **(F)** Scatterplot of pearson correlation of IP components and number of overlapping significant variants.

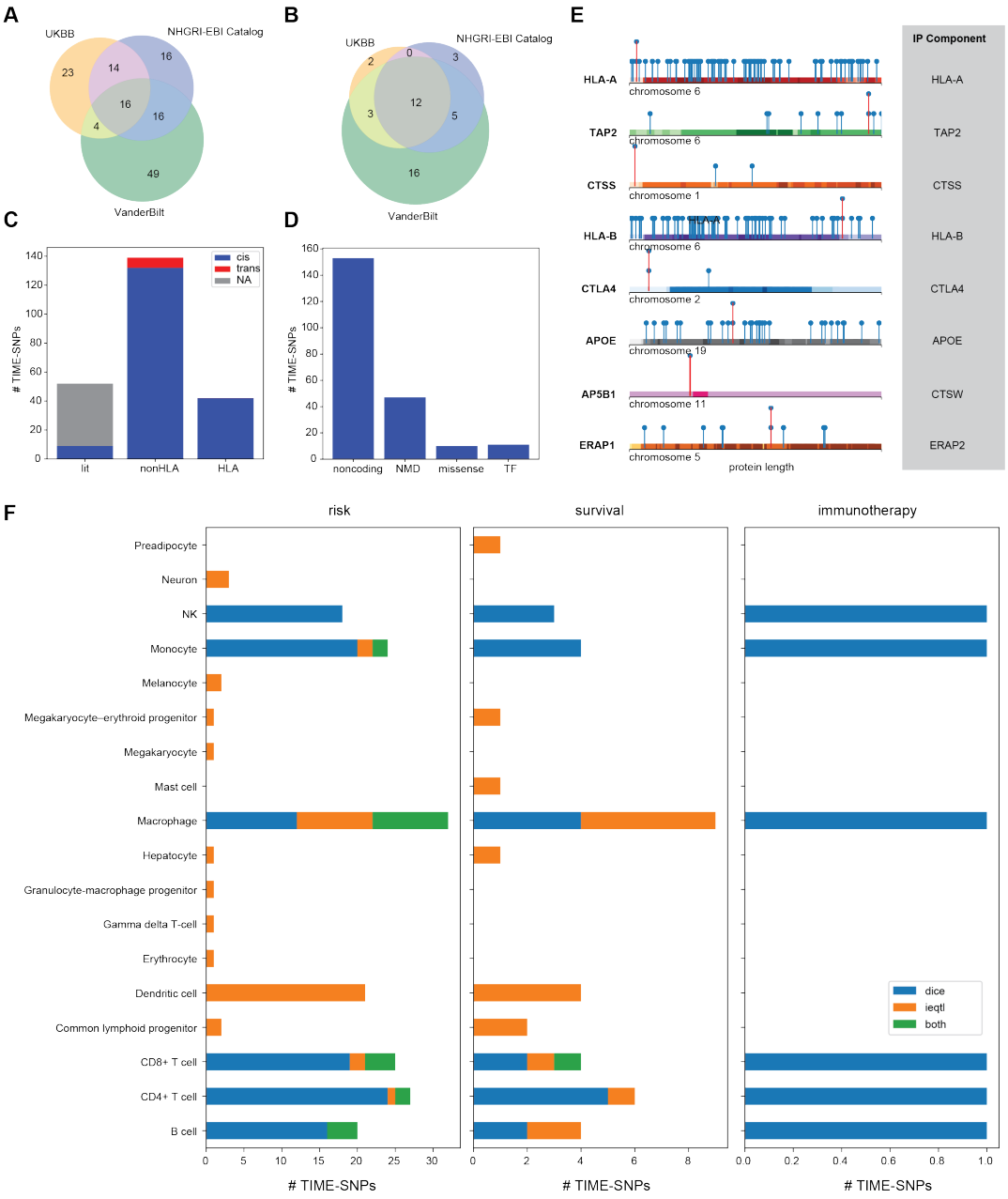

Figure S1.4 Identification and Characterization TIME-SNPs Related to Cancer Outcomes **(A)** Venn Diagram of overlap between TIME-SNPs implicated by UK Biobank PheWAS, Vanderbilt PheWAS catalog and NHGRI-EBI GWAS Catalog (LDtrait). **(B)** Venn Diagram of overlap between IP Components implicated by TIME-SNPs by UK Biobank PheWAS, Vanderbilt PheWAS catalog and NHGRI-EBI GWAS Catalog. **(C)** Barplot quantifying the number of cancer relevant TIME-SNPs which are *trans* (>1 MB from the TSS of the associated IP component) and *cis*. **(D)**. Barplot describing the effects of cancer relevant TIME-SNPs. **(E)** Location of missense TIME-SNPs (red) and known genetic variation (blue) in the coding sequence of affected IP components. The IP component with which the SNP was associated is shown on the right side. **(F)** Boxplot of number of cancer risk, survival and immunotherapy response TIME-SNPs which are cell-type eQTLs.

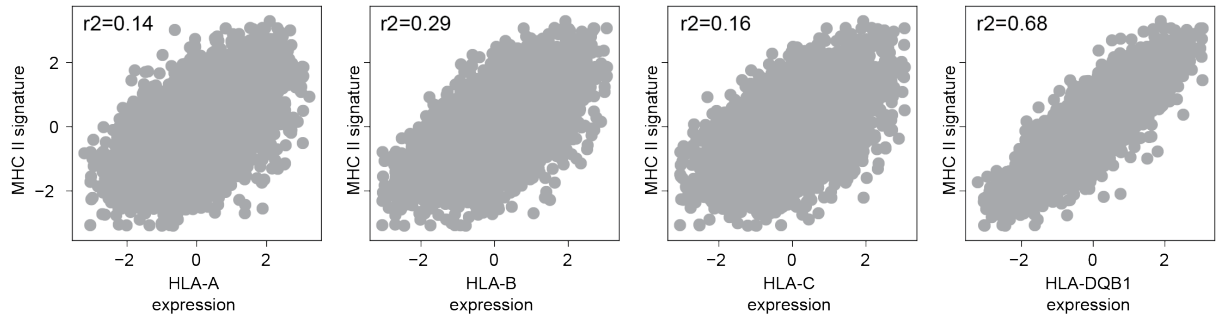

Figure S1.5 TIME-SNPs Underlying Antigen Presentation Stratify Melanoma and Prostate Cancer Risk Correlation between Sayaman et al. MHC Class II signature and HLA-A, HLA-B, HLA-C and HLA-DQB1 expression.

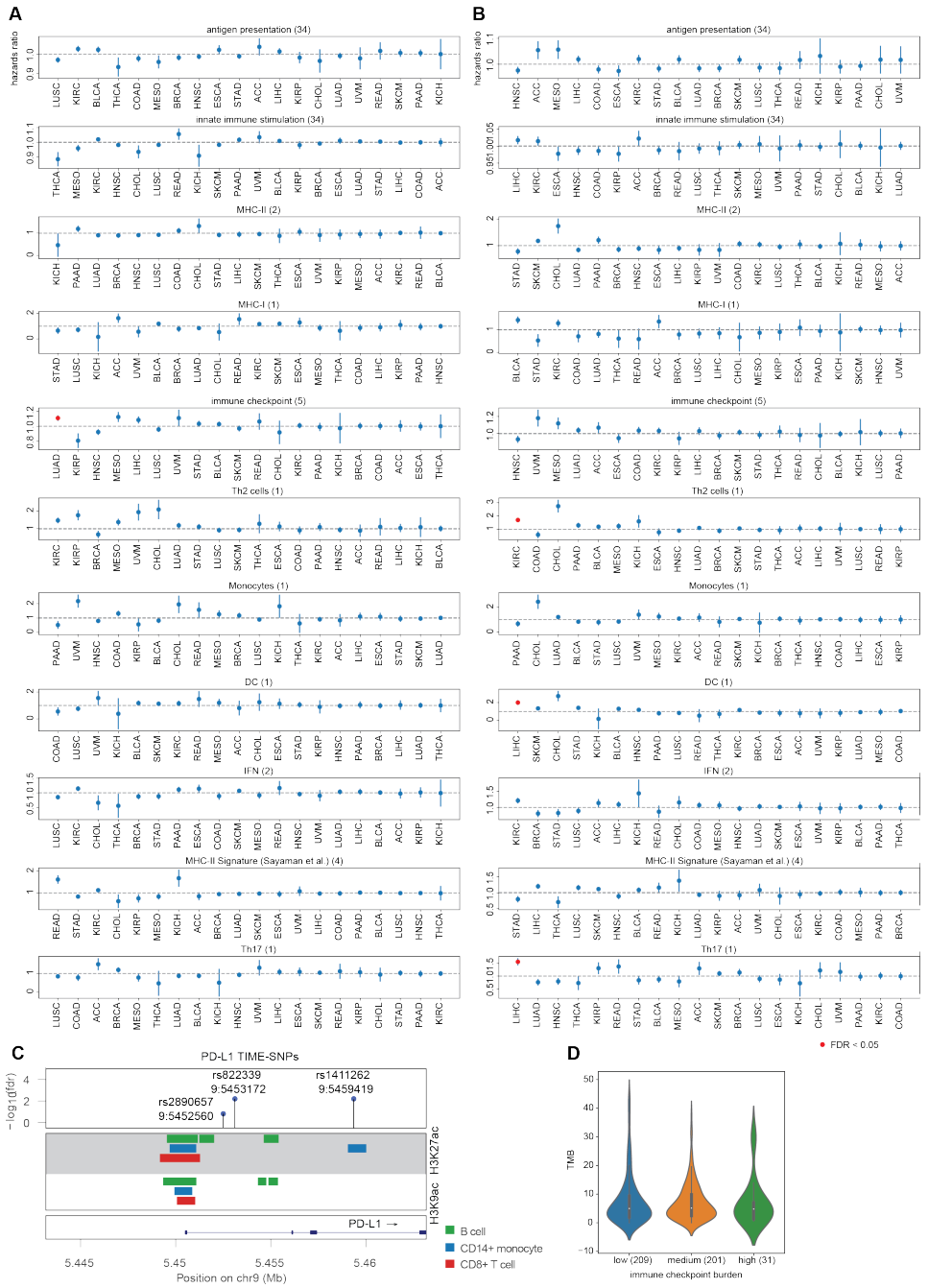

Figure S1.6 Variants Underlying Immune Evasion are Associated with Cancer Survival (A) Cox Proportional-Hazards results from association with burden score, if more than 1 variant association for IP category was detected, and overall survival for each TCGA cancer type. **(B)**. Cox Proportional-Hazards results from association with burden score, if more than 1 variant association for IP category was detected, and progression-free survival for each TCGA cancer type. **(C)** Location of PD-L1 variants implicated through literature review in relation to cell-type specific H3K27ac and H3K9ac marks. **(D)** Violin plot of immune checkpoint burden score and tumor mutational burden.

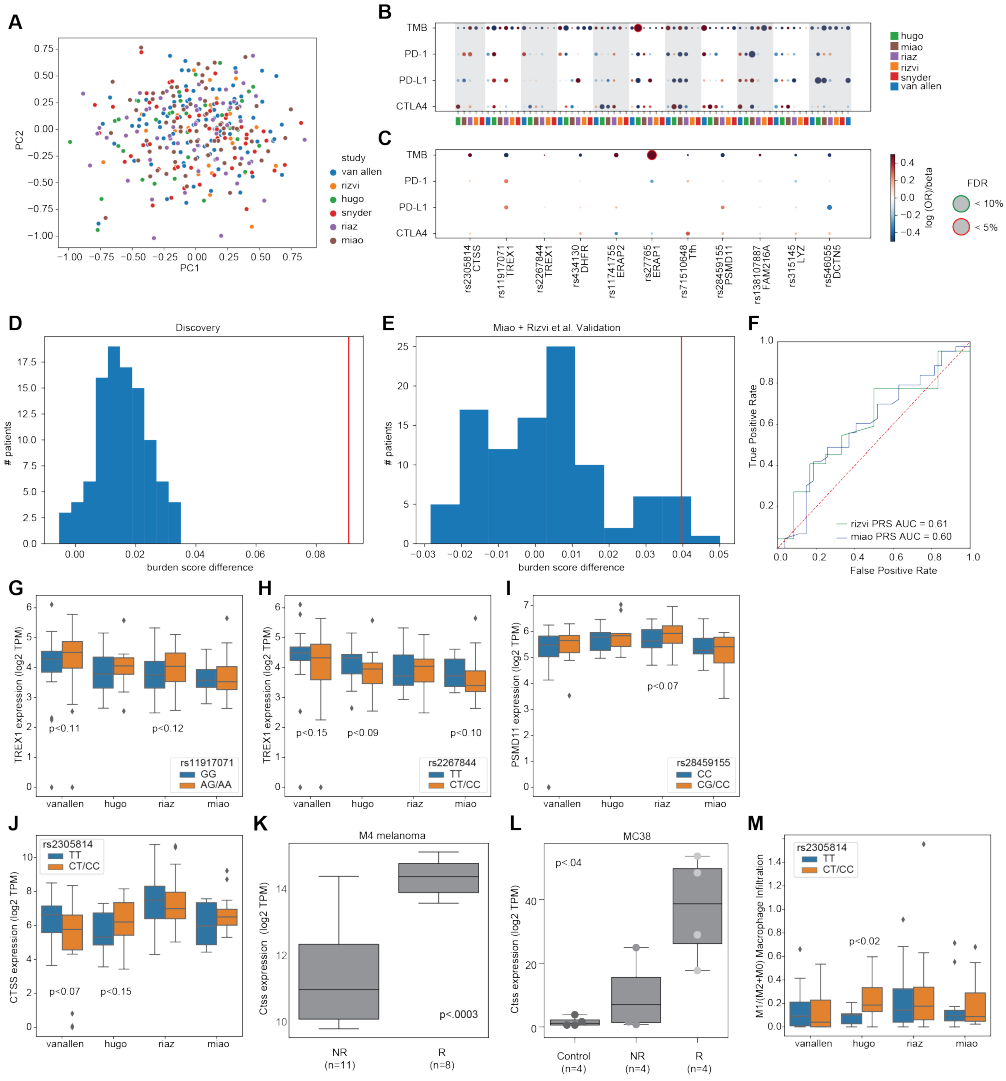

Figure S1.7 TIME-SNPs Implicate Targets for Modulating Immune Responses **(A)** PCA plot of 11 TIME-SNPs associated with immunotherapy response (FDR<0.5) **(B)** Grid plot of beta values of variant association with TMB, *CTLA-4, PD-1*, and *PD-L1* by each cohort. **(C)** Grid plot of beta values of variant association with TMB, *CTLA-4, PD-1*, and *PD-L1* controlling for cohort **(D)** Histogram of difference between average burden score for 100 random bootstrapping trials compared to actual difference based on 11 TIME-SNPs (red) in discovery cohort. **(E)** Histogram of difference between average burden score for 100 random bootstrapping trials compared to actual difference based on 11 TIME-SNPs (red) in Miao and Rizvi et al validation cohort. **(F)** ROC-AUC Curve Analysis for PRSice PRS scores trained on the discovery cohort and tested on Miao et al and Rizvi et al. **(G)** Boxplot of rs11917071 association with *TREX1* expression across ICB cohorts. **(H)** Boxplot of rs2267844 association with *TREX1* expression across ICB cohorts. **(I)** Boxplot of rs28459155 association with *PSMD11* expression across ICB cohorts. **(J)** Boxplot of rs2305814 association with *CTSS* expression across ICB cohorts. **(K)** *Ctss* expression in non-responder and responder of M4 melanoma mouse model. **(L)** *Ctss* expression in non-responder and responder of MC38 mouse model. **(M)** Boxplot of rs2305814 association with M1/M2+M0 macrophage infiltration across ICB cohorts.

1.1.9 Author Contributions

Original concept: Meghana Pagadala, Hannah Carter

Statistical analysis: Wesley Thompson, Chun Chieh Fan, Meghana Pagadala, Hannah Carter

PRS analysis: James Talwar, Meghana Pagadala, Timothy Sears

DICE analysis: Benjamin Schmiedel, Cristian Gonzalez-Colin, Pandurangan Vijayanand

UK Biobank Analysis: Steven Cao, Hyo Kim, Rany Salem

Mouse Validation: Victoria Wu, Silvio Gutkind, Chi-Ping Day, Eva Pérez-Guijarro, Glenn Merlino,

Cell-type eQTL analysis: Shervin Goudarzi, Divya Kirani

Manuscript writing: Meghana Pagadala, Hannah Carter, Pandurangan Vijayanand, Wesley Thompson, Maurizio Zanetti, Jill Mesirov, Sandip Patel, Olivier Harismendy, Chun Chieh Fan, Geral Morris, Chi-Ping Day, Andrea Castro

1.1.10 Acknowledgements

This work was supported by Emerging Leader Award from The Mark Foundation for Cancer Research, grant #18-022-ELA to HC, NIH grant R01CA220009 to HC and MZ, U01CA196406 to OH, and 1F30CA247168-01 and T32CA067754 to MP. UK Biobank data was retrieved under project ID 37671.

The results shown here are in large part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga and Genotype-Tissue Expression (GTEx) Project: https://gtexportal.org/home/. GTEx was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this manuscript were obtained from the GTEx Portal on 10/10/20.

This research has been conducted using the UK Biobank Resource under project ID 37671. Cancer risk validation cohorts were obtained from dbgap accessions: phs000187.v1.p1 and phs001125.v1.p1. Research support to collect data for phs000187.v1.p1 and develop an application to support phs000187.v1.p1 was provided by 3P50CA093459, 5P50CA097007, 5R01ES011740, and 5R01CA133996. Funding for the meta-analysis in phs001125.v1.p1 is provided by NIH grant U19CA148537. We would like to acknowledge the NCRN nurses and Consultants for their work in the UKGPCS study. We thank all the patients who took part in this study. This work was supported by Cancer Research UK (grant numbers C5047/A7357, C1287/A10118, C1287/A5260, C5047/A3354, C5047/A10692, C16913/A6135 and C16913/A6835).

For Rizvi *et al.* non-small cell lung cancer immunotherapy analysis, we used dbGaP data from accession phs000980.v1.p1. We thank the members of the Thoracic Oncology Service and

the Chan and Wolchok labs at MSKCC for helpful discussions. We thank the Immune Monitoring Core at MSKCC, including L. Caro, R. Ramsawak, and Z. Mu, for exceptional support with processing and banking peripheral blood lymphocytes. We thank P. Worrell and E. Brzostowski for help in identifying tumor specimens for analysis. We thank A. Viale for superb technical assistance. We thank D. Philips, M. van Buuren, and M. Toebes for help performing the combinatorial coding screens. The data presented in this paper are tabulated in the main paper and in the supplementary materials. This work was supported by the Geoffrey Beene Cancer Research Center (MDH, NAR, TAC, JDW, AS), the Society for Memorial Sloan Kettering Cancer Center (MDH), Lung Cancer Research Foundation (WL), Frederick Adler Chair Fund (TAC), The One Ball Matt Memorial Golf Tournament (EBG), Queen Wilhelmina Cancer Research Award (TNS), The STARR Foundation (TAC, JDW), the Ludwig Trust (JDW), and a Stand Up To Cancer-Cancer Research Institute Cancer Immunology Translational Cancer Research Grant (JDW, TNS, TAC). Stand Up To Cancer is a program of the Entertainment Industry Foundation administered by the American Association for Cancer Research. For Snyder et al. melanoma immunotherapy analysis, we used dbGaP data from accession phs001041.v1.p1. We thank Martin Miller at Memorial Sloan Kettering Cancer Center (MSKCC) for his assistance with the NetMHC server, Agnes Viale and Kety Huberman at the MSKCC Genomics Core, Annamalai Selvakumar and Alice Yeh at the MSKCC HLA typing laboratory for their technical assistance, and John Khoury for assistance in chart review. For Miao et al renal cell carcinoma immunotherapy analysis, we used dbGap data from accession phs001493.v2.p1. This study was supported by an AACR KureIt grant. Hugo *et al.* melanoma samples were acquired from SRA using accession numbers SRP067938 and SRP090294. Riaz *et al.* melanoma samples were acquired from SRA using

accession number SRP095809. For Van Allen *et al.* melanoma sample, data was acquired from dbgap accession phs000452.v2.p1.

Chapter 1, in full, includes reformatted reprint of the material as it appears in "Germline modifiers of the tumor immune microenvironment implicate drivers of cancer risk and immunotherapy response." in *biorxiv,* 2020 by M. Pagadala, V. H. Wu, E. Pérez-Guijarro, H. Kim, A. Castro, J. Talwar, C. Gonzalez-Colin, S. Cao, B. J. Schmiedel, R. M. Salem, G. P. Morris, O. Harismendy, S. P. Patel, J. P. Mesirov, M. Zanetti, C.-P. Day, C. C. Fan, W. K. Thompson, G. Merlino, J. Silvio Gutkind, P. Vijayanand, H. Carter. The dissertation author was a primary investigator and author of this paper.

1.1.11 References

1. M. R. Stratton, P. J. Campbell, P. Andrew Futreal, The cancer genome. *Nature*. **458** (2009), pp. 719–724.

2. B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz, K. W. Kinzler, Cancer Genome Landscapes. *Science*. **339** (2013), pp. 1546–1558.

3. D. Hanahan, R. A. Weinberg, The Hallmarks of Cancer. *Cell*. **100** (2000), pp. 57–70.

4. D. Hanahan, R. A. Weinberg, Hallmarks of Cancer: The Next Generation. *Cell*. **144** (2011), pp. 646–674.

5. S. I. Grivennikov, F. R. Greten, M. Karin, Immunity, inflammation, and cancer. *Cell*. **140**, 883–899 (2010).

6. G. P. Dunn, A. T. Bruce, H. Ikeda, L. J. Old, R. D. Schreiber, Cancer immunoediting: from immunosurveillance to tumor escape. *Nat. Immunol.* **3**, 991–998 (2002).

7. E. Borcoman, A. Nandikolla, G. Long, S. Goel, C. Le Tourneau, Patterns of Response and Progression to Immunotherapy. *American Society of Clinical Oncology Educational Book* (2018), pp. 169–178.

8. D. Miao, E. M. Van Allen, Genomic determinants of cancer immunotherapy. *Current Opinion in Immunology*. **41** (2016), pp. 32–38.

9. P. Geeleher, A. Nath, F. Wang, Z. Zhang, A. N. Barbeira, J. Fessler, R. L. Grossman, C. Seoighe, R. Stephanie Huang, Cancer expression quantitative trait loci (eQTLs) can be determined from heterogeneous tumor gene expression data by modeling variation in tumor purity. *Genome Biol.* **19**, 130 (2018).

10. S. Shahamatdar, M. X. He, M. A. Reyna, A. Gusev, S. H. AlDubayan, E. M. Van Allen, S. Ramachandran, Germline Features Associated with Immune Infiltration in Solid Tumors. *Cell Rep.* **30**, 2900–2908.e4 (2020).

11. R. W. Sayaman, M. Saad, V. Thorsson, W. Hendrickx, J. Roelands, Y. Mokrab, F. Farshidfar, T. Kirchhoff, R. F. Sweis, O. F. Bathe, E. Porta-Pardo, M. J. Campbell, C. Stretch, D. Hu, S. Huntsman, R. E. Graff, N. Syed, L. Radvanyi, S. Shelley, D. Wolf, F. M. Marincola, M. Ceccarelli, J. Galon, E. Ziv, D. Bedognetti, Germline genetic contribution to the immune landscape of cancer, , doi:10.1101/2020.01.30.926527.

12. Y. W. Lim, H. Chen-Harris, O. Mayba, S. Lianoglou, A. Wuster, T. Bhangale, Z. Khan, S. Mariathasan, A. Daemen, J. Reeder, P. M. Haverty, W. F. Forrest, M. Brauer, I. Mellman, M. L. Albert, Germline genetic polymorphisms influence tumor gene expression and immune cell infiltration. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E11701–E11710 (2018).

13. B. J. Schmiedel, D. Singh, A. Madrigal, A. G. Valdovino-Gonzalez, B. M. White, J.

Zapardiel-Gonzalo, B. Ha, G. Altay, J. A. Greenbaum, G. McVicker, G. Seumois, A. Rao, M. Kronenberg, B. Peters, P. Vijayanand, Impact of Genetic Polymorphisms on Human Immune Cell Gene Expression. *Cell*. **175** (2018), pp. 1701–1715.e16.

14. M. E. Hauberg, W. Zhang, C. Giambartolomei, O. Franzén, D. L. Morris, T. J. Vyse, A. Ruusalepp, CommonMind Consortium, P. Sklar, E. E. Schadt, J. L. M. Björkegren, P. Roussos, Large-Scale Identification of Common Trait and Disease Variants Affecting Gene Expression. *Am. J. Hum. Genet.* **101**, 157 (2017).

15. F. Hormozdiari, M. van de Bunt, A. V. Segrè, X. Li, J. W. J. Joo, M. Bilow, J. H. Sul, S. Sankararaman, B. Pasaniuc, E. Eskin, Colocalization of GWAS and eQTL Signals Detects Target Genes. *Am. J. Hum. Genet.* **99**, 1245–1260 (2016).

16. S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, P. C. Sham, PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*. **81** (2007), pp. 559–575.

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

18. J. Bryois, A. Buil, D. M. Evans, J. P. Kemp, S. B. Montgomery, D. F. Conrad, K. M. Ho, S. Ring, M. Hurles, P. Deloukas, G. D. Smith, E. T. Dermitzakis, Cis and Trans Effects of Human Genomic Variants on Gene Expression. *PLoS Genetics*. **10** (2014), p. e1004461.

19. C. Tian, B. S. Hromatka, A. K. Kiefer, N. Eriksson, S. M. Noble, J. Y. Tung, D. A. Hinds, Genome-wide association and HLA region fine-mapping studies identify susceptibility loci for multiple common infections. *Nat. Commun.* **8**, 599 (2017).

20. J. Knight, S. L. Spain, F. Capon, A. Hayday, F. O. Nestle, A. Clop, Wellcome Trust Case Control Consortium, Genetic Analysis of Psoriasis Consortium, I-chip for Psoriasis Consortium, J. N. Barker, M. E. Weale, R. C. Trembath, Conditional analysis identifies three novel major histocompatibility complex loci associated with psoriasis. *Hum. Mol. Genet.* **21**, 5185–5192 (2012).

21. V. R. C. Aguiar, C. Masotti, A. A. Camargo, D. Meyer, HLApers: HLA Typing and Quantification of Expression with Personalized Index. *Methods Mol. Biol.* **2120**, 101–112 (2020).

22. V. Thorsson, D. L. Gibbs, S. D. Brown, D. Wolf, D. S. Bortone, T.-H. Ou Yang, E. Porta-Pardo, G. F. Gao, C. L. Plaisier, J. A. Eddy, E. Ziv, A. C. Culhane, E. O. Paull, I. K. A. Sivakumar, A. J. Gentles, R. Malhotra, F. Farshidfar, A. Colaprico, J. S. Parker, L. E. Mose, N. S. Vo, J. Liu, Y. Liu, J. Rader, V. Dhankani, S. M. Reynolds, R. Bowlby, A. Califano, A. D. Cherniack, D. Anastassiou, D. Bedognetti, Y. Mokrab, A. M. Newman, A. Rao, K. Chen, A. Krasnitz, H. Hu, T. M. Malta, H. Noushmehr, C. S. Pedamallu, S. Bullman, A. I. Ojesina, A. Lamb, W. Zhou, H. Shen, T. K. Choueiri, J. N. Weinstein, J. Guinney, J. Saltz, R. A. Holt, C. S. Rabkin, Cancer Genome Atlas Research Network, A. J. Lazar, J. S. Serody, E. G. Demicco, M. L. Disis, B. G. Vincent, I. Shmulevich, The Immune Landscape of Cancer.

Immunity. **51**, 411–412 (2019).

23. B. N. Ostendorf, J. Bilanovic, N. Adaku, K. N. Tafreshian, B. Tavora, R. D. Vaughan, S. F. Tavazoie, Common germline variants of the human APOE gene modulate melanoma progression and survival. *Nat. Med.* **26**, 1048–1053 (2020).

24. Y. Zhang, M. Manjunath, J. Yan, B. A. Baur, S. Zhang, S. Roy, J. S. Song, The Cancer-Associated Genetic Variant Rs3903072 Modulates Immune Cells in the Tumor Microenvironment. *Front. Genet.* **10**, 754 (2019).

25. P. Queirolo, B. Dozin, A. Morabito, B. Banelli, P. Piccioli, C. Fava, C. Leo, R. Carosio, S. Laurent, V. Fontana, P. F. Ferrucci, C. Martinoli, E. Cocorocchio, A. Battaglia, P. A. Ascierto, M. Capone, E. Simeone, F. De Galitiis, E. Pagani, G. C. Antonini Cappellini, P. Marchetti, M. Guida, S. Tommasi, M. Mandalà, B. Merelli, P. Quaglino, P. Fava, M. Guidoboni, M. Romani, F. Spagnolo, M. P. Pistillo, Association of CTLA-4 Gene Variants with Response to Therapy and Long-term Survival in Metastatic Melanoma Patients Treated with Ipilimumab: An Italian Melanoma Intergroup Study. *Front. Immunol.* **8**, 386 (2017).

26. H. Yoshida, T. Nomizo, H. Ozasa, T. Tsuji, T. Funazo, Y. Yasuda, H. Ajimizu, M. Yamazoe, K. Kuninaga, T. Ogimoto, K. Hosoya, R. Itotani, Y. Sakamori, Y. H. Kim, T. Hirai, PD-L1 polymorphisms predict survival outcomes in advanced non-small-cell lung cancer patients treated with PD-1 blockade. *Eur. J. Cancer*. **144**, 317–325 (2021).

27. A. Kula, M. Dawidowicz, P. Kiczmer, A. Prawdzic Seńkowska, E. Świętochowska, The role of genetic polymorphism within PD-L1 gene in cancer. Review. *Exp. Mol. Pathol.* **116**, 104494 (2020).

28. A. Salmaninejad, V. Khoramshahi, A. Azani, E. Soltaninejad, S. Aslani, M. R. Zamani, M. Zal, A. Nesaei, S. M. Hosseini, PD-1 and cancer: molecular mechanisms and polymorphisms. *Immunogenetics*. **70**, 73–86 (2018).

29. H. Sasaki, T. Tatemaysu, K. Okuda, S. Moriyama, M. Yano, Y. Fujii, PD-1 gene promoter polymorphisms correlate with a poor prognosis in non-small cell lung cancer. *Molecular and Clinical Oncology*. **2**, 1035–1042 (2014).

30. W. Tang, Y. Chen, S. Chen, B. Sun, H. Gu, M. Kang, Programmed death-1 (PD-1) polymorphism is associated with gastric cardia adenocarcinoma. *Int. J. Clin. Exp. Med.* **8**, 8086–8093 (2015).

31. D. Bedognetti, T. L. Spivey, Y. Zhao, L. Uccellini, S. Tomei, M. E. Dudley, M. L. Ascierto, V. De Giorgi, Q. Liu, L. G. Delogu, M. Sommariva, M. R. Sertoli, R. Simon, E. Wang, S. A. Rosenberg, F. M. Marincola, CXCR3/CCR5 pathways in metastatic melanoma patients treated with adoptive therapy and interleukin-2. *Br. J. Cancer*. **109**, 2412–2423 (2013).

32. L. Uccellini, V. De Giorgi, Y. Zhao, B. Tumaini, N. Erdenebileg, M. E. Dudley, S. Tomei, D. Bedognetti, M. L. Ascierto, Q. Liu, R. Simon, L. Kottyan, K. M. Kaufman, J. B. Harley, E. Wang, S. A. Rosenberg, F. M. Marincola, IRF5 gene polymorphisms in melanoma. *J. Transl. Med.* **10**, 170 (2012).

33. D. Kogan, A. Grabner, C. Yanucil, C. Faul, V. K. Ulaganathan, STAT3-enhancing germline mutations contribute to tumor-extrinsic immune evasion. *J. Clin. Invest.* **128**, 1867–1872 (2018).

34. Germline genetic contribution to the immune landscape of cancer. *Immunity*. **54**, 367–386.e8 (2021).

35. K. Michailidou, S. Lindström, J. Dennis, J. Beesley, S. Hui, S. Kar, A. Lemaçon, P. Soucy, D. Glubb, A. Rostamianfar, M. K. Bolla, Q. Wang, J. Tyrer, E. Dicks, A. Lee, Z. Wang, J. Allen, R. Keeman, U. Eilber, J. D. French, X. Qing Chen, L. Fachal, K. McCue, A. E. McCart Reed, M. Ghoussaini, J. S. Carroll, X. Jiang, H. Finucane, M. Adams, M. A. Adank, H. Ahsan, K. Aittomäki, H. Anton-Culver, N. N. Antonenkova, V. Arndt, K. J. Aronson, B. Arun, P. L. Auer, F. Bacot, M. Barrdahl, C. Baynes, M. W. Beckmann, S. Behrens, J. Benitez, M. Bermisheva, L. Bernstein, C. Blomqvist, N. V. Bogdanova, S. E. Bojesen, B. Bonanni, A.-L. Børresen-Dale, J. S. Brand, H. Brauch, P. Brennan, H. Brenner, L. Brinton, P. Broberg, I. W. Brock, A. Broeks, A. Brooks-Wilson, S. Y. Brucker, T. Brüning, B. Burwinkel, K. Butterbach, Q. Cai, H. Cai, T. Caldés, F. Canzian, A. Carracedo, B. D. Carter, J. E. Castelao, T. L. Chan, T.-Y. David Cheng, K. Seng Chia, J.-Y. Choi, H. Christiansen, C. L. Clarke, NBCS Collaborators, M. Collée, D. M. Conroy, E. Cordina-Duverger, S. Cornelissen, D. G. Cox, A. Cox, S. S. Cross, J. M. Cunningham, K. Czene, M. B. Daly, P. Devilee, K. F. Doheny, T. Dörk, I. Dos-Santos-Silva, M. Dumont, L. Durcan, M. Dwek, D. M. Eccles, A. B. Ekici, A. H. Eliassen, C. Ellberg, M. Elvira, C. Engel, M. Eriksson, P. A. Fasching, J. Figueroa, D. Flesch-Janys, O. Fletcher, H. Flyger, L. Fritschi, V. Gaborieau, M. Gabrielson, M. Gago-Dominguez, Y.-T. Gao, S. M. Gapstur, J. A. García-Sáenz, M. M. Gaudet, V. Georgoulias, G. G. Giles, G. Glendon, M. S. Goldberg, D. E. Goldgar, A. González-Neira, G. I. Grenaker Alnæs, M. Grip, J. Gronwald, A. Grundy, P. Guénel, L. Haeberle, E. Hahnen, C. A. Haiman, N. Håkansson, U. Hamann, N. Hamel, S. Hankinson, P. Harrington, S. N. Hart, J. M. Hartikainen, M. Hartman, A. Hein, J. Heyworth, B. Hicks, P. Hillemanns, D. N. Ho, A. Hollestelle, M. J. Hooning, R. N. Hoover, J. L. Hopper, M.-F. Hou, C.-N. Hsiung, G. Huang, K. Humphreys, J. Ishiguro, H. Ito, M. Iwasaki, H. Iwata, A. Jakubowska, W. Janni, E. M. John, N. Johnson, K. Jones, M. Jones, A. Jukkola-Vuorinen, R. Kaaks, M. Kabisch, K. Kaczmarek, D. Kang, Y. Kasuga, M. J. Kerin, S. Khan, E. Khusnutdinova, J. I. Kiiski, S.-W. Kim, J. A. Knight, V.-M. Kosma, V. N. Kristensen, U. Krüger, A. Kwong, D. Lambrechts, L. Le Marchand, E. Lee, M. H. Lee, J. W. Lee, C. Neng Lee, F. Lejbkowicz, J. Li, J. Lilyquist, A. Lindblom, J. Lissowska, W.-Y. Lo, S. Loibl, J. Long, A. Lophatananon, J. Lubinski, C. Luccarini, M. P. Lux, E. S. K. Ma, R. J. MacInnis, T. Maishman, E. Makalic, K. E. Malone, I. M. Kostovska, A. Mannermaa, S. Manoukian, J. E. Manson, S. Margolin, S. Mariapun, M. E. Martinez, K. Matsuo, D. Mavroudis, J. McKay, C. McLean, H. Meijers-Heijboer, A. Meindl, P. Menéndez, U. Menon, J. Meyer, H. Miao, N. Miller, N. A. M. Taib, K. Muir, A. M. Mulligan, C. Mulot, S. L. Neuhausen, H. Nevanlinna, P. Neven, S. F. Nielsen, D.-Y. Noh, B. G. Nordestgaard, A. Norman, O. I. Olopade, J. E. Olson, H. Olsson, C. Olswold, N. Orr, V. S. Pankratz, S. K. Park, T.-W. Park-Simon, R. Lloyd, J. I. A. Perez, P. Peterlongo, J. Peto, K.-A. Phillips, M. Pinchev, D. Plaseska-Karanfilska, R. Prentice, N. Presneau, D. Prokofyeva, E. Pugh, K. Pylkäs, B. Rack, P. Radice, N. Rahman, G. Rennert, H. S. Rennert, V. Rhenius, A. Romero, J. Romm, K. J. Ruddy, T. Rüdiger, A. Rudolph, M. Ruebner, E. J. T. Rutgers, E. Saloustros, D. P. Sandler, S. Sangrajrang, E. J. Sawyer, D. F. Schmidt, R. K. Schmutzler, A. Schneeweiss, M. J. Schoemaker, F. Schumacher, P. Schürmann, R. J. Scott,

C. Scott, S. Seal, C. Seynaeve, M. Shah, P. Sharma, C.-Y. Shen, G. Sheng, M. E. Sherman, M. J. Shrubsole, X.-O. Shu, A. Smeets, C. Sohn, M. C. Southey, J. J. Spinelli, C. Stegmaier, S. Stewart-Brown, J. Stone, D. O. Stram, H. Surowy, A. Swerdlow, R. Tamimi, J. A. Taylor, M. Tengström, S. H. Teo, M. Beth Terry, D. C. Tessier, S. Thanasitthichai, K. Thöne, R. A. E. M. Tollenaar, I. Tomlinson, L. Tong, D. Torres, T. Truong, C.-C. Tseng, S. Tsugane, H.-U. Ulmer, G. Ursin, M. Untch, C. Vachon, C. J. van Asperen, D. Van Den Berg, A. M. W. van den Ouweland, L. van der Kolk, R. B. van der Luijt, D. Vincent, J. Vollenweider, Q. Waisfisz, S. Wang-Gohrke, C. R. Weinberg, C. Wendt, A. S. Whittemore, H. Wildiers, W. Willett, R. Winqvist, A. Wolk, A. H. Wu, L. Xia, T. Yamaji, X. R. Yang, C. Har Yip, K.-Y. Yoo, J.-C. Yu, W. Zheng, Y. Zheng, B. Zhu, A. Ziogas, E. Ziv, ABCTB Investigators, ConFab/AOCS Investigators, S. R. Lakhani, A. C. Antoniou, A. Droit, I. L. Andrulis, C. I. Amos, F. J. Couch, P. D. P. Pharoah, J. Chang-Claude, P. Hall, D. J. Hunter, R. L. Milne, M. García-Closas, M. K. Schmidt, S. J. Chanock, A. M. Dunning, S. L. Edwards, G. D. Bader, G. Chenevix-Trench, J. Simard, P. Kraft, D. F. Easton, Association analysis identifies 65 new breast cancer risk loci. *Nature*. **551**, 92–94 (2017).

36. M. J. Machiela, S. J. Chanock, LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants: Fig. 1. *Bioinformatics*. **31** (2015), pp. 3555–3557.

37. T. A. Myers, S. J. Chanock, M. J. Machiela, LDlinkR: An R Package for Rapidly Calculating Linkage Disequilibrium Statistics in Diverse Populations. *Frontiers in Genetics*. **11** (2020), , doi:10.3389/fgene.2020.00157.

38. J. H. Karnes, L. Bastarache, C. M. Shaffer, S. Gaudieri, Y. Xu, A. M. Glazer, J. D. Mosley, S. Zhao, S. Raychaudhuri, S. Mallal, Z. Ye, J. G. Mayer, M. H. Brilliant, S. J. Hebbring, D. M. Roden, E. J. Phillips, J. C. Denny, Phenome-wide scanning identifies multiple diseases and disease severity phenotypes associated with HLA variants. *Sci. Transl. Med.* **9** (2017), doi:10.1126/scitranslmed.aai8708.

39. C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray, B. Liu, P. Matthews, G. Ong, J. Pell, A. Silman, A. Young, T. Sprosen, T. Peakman, R. Collins, UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).

40. D. M. Wolf, M. E. Lenburg, C. Yau, A. Boudreau, L. J. van 't Veer, Gene co-expression modules as clinically relevant hallmarks of breast cancer diversity. *PLoS One*. **9**, e88309 (2014).

41. W. Hugo, J. M. Zaretsky, L. Sun, C. Song, B. H. Moreno, S. Hu-Lieskovan, B. Berent-Maoz, J. Pang, B. Chmielowski, G. Cherry, E. Seja, S. Lomeli, X. Kong, M. C. Kelley, J. A. Sosman, D. B. Johnson, A. Ribas, R. S. Lo, Genomic and Transcriptomic Features of Response to Anti-PD-1 Therapy in Metastatic Melanoma. *Cell*. **168**, 542 (2017).

42. E. M. Van Allen, D. Miao, B. Schilling, S. A. Shukla, C. Blank, L. Zimmer, A. Sucker, U. Hillen, M. H. G. Foppen, S. M. Goldinger, J. Utikal, J. C. Hassel, B. Weide, K. C. Kaehler, C. Loquai, P. Mohr, R. Gutzmer, R. Dummer, S. Gabriel, C. J. Wu, D. Schadendorf, L. A.

Garraway, Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science*. **350**, 207–211 (2015).

43. N. A. Rizvi, M. D. Hellmann, A. Snyder, P. Kvistborg, V. Makarov, J. J. Havel, W. Lee, J. Yuan, P. Wong, T. S. Ho, M. L. Miller, N. Rekhtman, A. L. Moreira, F. Ibrahim, C. Bruggeman, B. Gasmi, R. Zappasodi, Y. Maeda, C. Sander, E. B. Garon, T. Merghoub, J. D. Wolchok, T. N. Schumacher, T. A. Chan, Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science*. **348**, 124–128 (2015).

44. N. Riaz, J. J. Havel, V. Makarov, A. Desrichard, W. J. Urba, J. S. Sims, F. S. Hodi, S. Martín-Algarra, R. Mandal, W. H. Sharfman, S. Bhatia, W.-J. Hwu, T. F. Gajewski, C. L. Slingluff Jr, D. Chowell, S. M. Kendall, H. Chang, R. Shah, F. Kuo, L. G. T. Morris, J.-W. Sidhom, J. P. Schneck, C. E. Horak, N. Weinhold, T. A. Chan, Tumor and Microenvironment Evolution during Immunotherapy with Nivolumab. *Cell*. **171**, 934–949.e16 (2017).

45. A. Snyder, V. Makarov, T. Merghoub, J. Yuan, J. M. Zaretsky, A. Desrichard, L. A. Walsh, M. A. Postow, P. Wong, T. S. Ho, T. J. Hollmann, C. Bruggeman, K. Kannan, Y. Li, C. Elipenahli, C. Liu, C. T. Harbison, L. Wang, A. Ribas, J. D. Wolchok, T. A. Chan, Genetic basis for clinical response to CTLA-4 blockade in melanoma. *N. Engl. J. Med.* **371**, 2189–2199 (2014).

46. P. I. W. de Bakker, M. A. R. Ferreira, X. Jia, B. M. Neale, S. Raychaudhuri, B. F. Voight, Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum. Mol. Genet.* **17**, R122–8 (2008).

47. E. Kreiner-Møller, C. Medina-Gomez, A. G. Uitterlinden, F. Rivadeneira, K. Estrada, Improving accuracy of rare variant imputation with a two-step imputation approach. *Eur. J. Hum. Genet.* **23**, 395–400 (2015).

48. C. J. Willer, Y. Li, G. R. Abecasis, METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*. **26**, 2190–2191 (2010).

49. A. J. Bannister, T. Kouzarides, Regulation of chromatin by histone modifications. *Cell Res.* **21**, 381–395 (2011).

50. G. McVicker, B. van de Geijn, J. F. Degner, C. E. Cain, N. E. Banovich, A. Raj, N. Lewellen, M. Myrthil, Y. Gilad, J. K. Pritchard, Identification of genetic variants that affect histone modifications in human cells. *Science*. **342**, 747–749 (2013).

51. M. P. Creyghton, A. W. Cheng, G. G. Welstead, T. Kooistra, B. W. Carey, E. J. Steine, J. Hanna, M. A. Lodato, G. M. Frampton, P. A. Sharp, L. A. Boyer, R. A. Young, R. Jaenisch, Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 21931–21936 (2010).

52. G. Liang, J. C. Y. Lin, V. Wei, C. Yoo, J. C. Cheng, C. T. Nguyen, D. J. Weisenberger, G. Egger, D. Takai, F. A. Gonzales, P. A. Jones, Distinct localization of histone H3 acetylation and H3-K4 methylation to the transcription start sites in the human genome. *Proc. Natl. Acad.*

Sci. U. S. A. **101**, 7357–7362 (2004).

53. V. Sasidharan Nair, H. El Salhat, R. Z. Taha, A. John, B. R. Ali, E. Elkord, DNA methylation and repressive H3K9 and H3K27 trimethylation in the promoter regions of PD-1, CTLA-4, TIM-3, LAG-3, TIGIT, and PD-L1 genes in human primary breast cancer. *Clin. Epigenetics*. **10**, 78 (2018).

54. Y. Zhou, I. N. Bastian, M. D. Long, M. Dow, W. Li, T. Liu, R. K. Ngu, L. Antonucci, J. Y. Huang, Q. T. Phung, X.-H. Zhao, S. Banerjee, X.-J. Lin, H. Wang, B. Dang, S. Choi, D. Karin, H. Su, M. H. Ellisman, C. Jamieson, M. Bosenberg, Z. Cheng, J. Haybaeck, L. Kenner, K. M. Fisch, R. Bourgon, G. Hernandez, J. R. Lill, S. Liu, H. Carter, I. Mellman, M. Karin, S. Shalapour, Activation of NF-κB and p300/CBP potentiates cancer chemoimmunotherapy through induction of MHC-I antigen presentation. *Proc. Natl. Acad. Sci. U. S. A.* **118** (2021), doi:10.1073/pnas.2025840118.

55. R. Krupar, C. Watermann, C. Idel, J. Ribbat-Idel, A. Offermann, H. Pasternack, J. Kirfel, A. G. Sikora, S. Perner, In silico analysis reveals EP300 as a panCancer inhibitor of anti-tumor immune response via metabolic modulation. *Sci. Rep.* **10**, 9389 (2020).

56. D. R. Principe, R. Xiong, Y. Li, T. N. D. Pham, S. D. Kamath, O. Dubrovskyi, K. Ratia, F. Huang, J. Zhao, Z. Shen, D. Thummuri, Z. Daohong, P. W. Underwood, J. Trevino, H. G. Munshi, G. R. J. Thatcher, A. Rana, XP-524 is a dual-BET/EP300 inhibitor that represses oncogenic KRAS and potentiates immune checkpoint inhibition in pancreatic cancer. *Proc. Natl. Acad. Sci. U. S. A.* **119** (2022), doi:10.1073/pnas.2116764119.

57. C. Li, Z. Liu, L. E. Wang, J. E. Gershenwald, J. E. Lee, V. G. Prieto, M. Duvic, E. A. Grimm, Q. Wei, Haplotype and genotypes of the VDR gene and cutaneous melanoma risk in non-Hispanic whites in Texas: a case-control study. *Int. J. Cancer*. **122**, 2077–2084 (2008).

58. A. A. Al Olama, Z. Kote-Jarai, S. I. Berndt, D. V. Conti, F. Schumacher, Y. Han, S. Benlloch, D. J. Hazelett, Z. Wang, E. Saunders, D. Leongamornlert, S. Lindstrom, S. Jugurnauth-Little, T. Dadaev, M. Tymrakiewicz, D. O. Stram, K. Rand, P. Wan, A. Stram, X. Sheng, L. C. Pooler, K. Park, L. Xia, J. Tyrer, L. N. Kolonel, L. Le Marchand, R. N. Hoover, M. J. Machiela, M. Yeager, L. Burdette, C. C. Chung, A. Hutchinson, K. Yu, C. Goh, M. Ahmed, K. Govindasami, M. Guy, T. L. J. Tammela, A. Auvinen, T. Wahlfors, J. Schleutker, T. Visakorpi, K. A. Leinonen, J. Xu, M. Aly, J. Donovan, R. C. Travis, T. J. Key, A. Siddiq, F. Canzian, K.-T. Khaw, A. Takahashi, M. Kubo, P. Pharoah, N. Pashayan, M. Weischer, B. G. Nordestgaard, S. F. Nielsen, P. Klarskov, M. A. Røder, P. Iversen, S. N. Thibodeau, S. K. McDonnell, D. J. Schaid, J. L. Stanford, S. Kolb, S. Holt, B. Knudsen, A. H. Coll, S. M. Gapstur, W. R. Diver, V. L. Stevens, C. Maier, M. Luedeke, K. Herkommer, A. E. Rinckleb, S. S. Strom, C. Pettaway, E. D. Yeboah, Y. Tettey, R. B. Biritwum, A. A. Adjei, E. Tay, A. Truelove, S. Niwa, A. P. Chokkalingam, L. Cannon-Albright, C. Cybulski, D. Wokołorczyk, W. Kluźniak, J. Park, T. Sellers, H.-Y. Lin, W. B. Isaacs, A. W. Partin, H. Brenner, A. K. Dieffenbach, C. Stegmaier, C. Chen, E. L. Giovannucci, J. Ma, M. Stampfer, K. L. Penney, L. Mucci, E. M. John, S. A. Ingles, R. A. Kittles, A. B. Murphy, H. Pandha, A. Michael, A. M. Kierzek, W. Blot, L. B. Signorello, W. Zheng, D. Albanes, J. Virtamo, S. Weinstein, B. Nemesure, J. Carpten, C. Leske, S.-Y. Wu, A. Hennis, A. S. Kibel, B. A. Rybicki, C. Neslund-

Dudas, A. W. Hsing, L. Chu, P. J. Goodman, E. A. Klein, S. L. Zheng, J. Batra, J. Clements, A. Spurdle, M. R. Teixeira, P. Paulo, S. Maia, C. Slavov, R. Kaneva, V. Mitev, J. S. Witte, G. Casey, E. M. Gillanders, D. Seminara, E. Riboli, F. C. Hamdy, G. A. Coetzee, Q. Li, M. L. Freedman, D. J. Hunter, K. Muir, H. Gronberg, D. E. Neal, M. Southey, G. G. Giles, G. Severi, Breast and Prostate Cancer Cohort Consortium (BPC3), PRACTICAL (Prostate Cancer Association Group to Investigate Cancer-Associated Alterations in the Genome) Consortium, COGS (Collaborative Oncological Gene-environment Study) Consortium, GAME-ON/ELLIPSE Consortium, M. B. Cook, H. Nakagawa, F. Wiklund, P. Kraft, S. J. Chanock, B. E. Henderson, D. F. Easton, R. A. Eeles, C. A. Haiman, A meta-analysis of 87,040 individuals identifies 23 new susceptibility loci for prostate cancer. *Nat. Genet.* **46**, 1103–1109 (2014).

59. R. Liao, J. Sun, H. Wu, Y. Yi, J.-X. Wang, H.-W. He, X.-Y. Cai, J. Zhou, Y.-F. Cheng, J. Fan, S.-J. Qiu, High expression of IL-17 and IL-17RE associate with poor prognosis of hepatocellular carcinoma. *J. Exp. Clin. Cancer Res.* **32**, 3 (2013).

60. J.-P. Zhang, J. Yan, J. Xu, X.-H. Pang, M.-S. Chen, L. Li, C. Wu, S.-P. Li, L. Zheng, Increased intratumoral IL-17-producing cells correlate with poor survival in hepatocellular carcinoma patients. *J. Hepatol.* **50**, 980–989 (2009).

61. W. O. Hemphill, S. R. Simpson, M. Liu, F. R. Salsbury Jr, T. Hollis, J. M. Grayson, F. W. Perrino, TREX1 as a Novel Immunotherapeutic Target. *Front. Immunol.* **12**, 660184 (2021).

62. D. Chowdhury, P. J. Beresford, P. Zhu, D. Zhang, J.-S. Sung, B. Demple, F. W. Perrino, J. Lieberman, The exonuclease TREX1 is in the SET complex and acts in concert with NM23-H1 to degrade DNA during granzyme A-mediated cell death. *Mol. Cell*. **23**, 133–142 (2006).

63. L. Wang, L. Zhao, G. Wei, D. Saur, B. Seidler, J. Wang, C. Wang, T. Qi, Homoharringtonine could induce quick protein synthesis of PSMD11 through activating MEK1/ERK1/2 signaling pathway in pancreatic cancer cells. *J. Cell. Biochem.* **119**, 6644–6656 (2018).

64. D. Bararia, J. A. Hildebrand, S. Stolz, S. Haebe, S. Alig, C. P. Trevisani, F. Osorio-Barrios, M. D. Bartoschek, M. Mentz, A. Pastore, E. Gaitzsch, M. Heide, V. Jurinovic, K. Rautter, J. Gunawardana, M. B. Sabdia, M. Szczepanowski, J. Richter, W. Klapper, A. Louissaint Jr, C. Ludwig, S. Bultmann, H. Leonhardt, S. Eustermann, K.-P. Hopfner, W. Hiddemann, M. von Bergwelt-Baildon, C. Steidl, R. Kridel, J. W. D. Tobin, M. K. Gandhi, D. M. Weinstock, M. Schmidt-Supprian, M. B. Sárosi, M. Rudelius, V. Passerini, J. Mautner, O. Weigert, Cathepsin S Alterations Induce a Tumor-Promoting Immune Microenvironment in Follicular Lymphoma. *Cell Rep.* **31**, 107522 (2020).

65. K. Chamoto, R. Hatae, T. Honjo, Current issues and perspectives in PD-1 blockade cancer immunotherapy. *Int. J. Clin. Oncol.* **25**, 790–800 (2020).

66. R. F. Sweis, J. J. Luke, Mechanistic and pharmacologic insights on immune checkpoint inhibitors. *Pharmacol. Res.* **120**, 1–9 (2017).

67. Y. Ishida, Y. Agata, K. Shibahara, T. Honjo, Induced expression of PD-1, a novel member of the immunoglobulin gene superfamily, upon programmed cell death. *EMBO J.* **11**, 3887–

3895 (1992).

68. D. R. Leach, M. F. Krummel, J. P. Allison, Enhancement of antitumor immunity by CTLA-4 blockade. *Science*. **271**, 1734–1736 (1996).

69. E. Dheilly, E. Battistello, N. Katanayeva, S. Sungalee, J. Michaux, G. Duns, S. Wehrle, J. Sordet-Dessimoz, M. Mina, J. Racle, P. Farinha, G. Coukos, D. Gfeller, A. Mottok, R. Kridel, B. E. Correia, C. Steidl, M. Bassani-Sternberg, G. Ciriello, V. Zoete, E. Oricchio, Cathepsin S Regulates Antigen Processing and T Cell Activity in Non-Hodgkin Lymphoma. *Cancer Cell*. **37** (2020), pp. 674–689.e12.

70. C. Riether, A. F. Ochsenbein, Genetic Alterations Impact Immune Microenvironment Interactions in Follicular Lymphoma. *Cancer Cell*. **37** (2020), pp. 621–622.

71. F. R. Greten, S. I. Grivennikov, Inflammation and Cancer: Triggers, Mechanisms, and Consequences. *Immunity*. **51**, 27–41 (2019).

72. V. S. Pelly, A. Moeini, L. M. Roelofsen, E. Bonavita, C. R. Bell, C. Hutton, A. Blanco-Gomez, A. Banyard, C. P. Bromley, E. Flanagan, S.-C. Chiang, C. Jørgensen, T. N. Schumacher, D. S. Thommen, S. Zelenay, Anti-Inflammatory Drugs Remodel the Tumor Immune Environment to Enhance Immune Checkpoint Blockade Efficacy. *Cancer Discov.* **11**, 2602–2619 (2021).

73. D. Chowell, L. G. T. Morris, C. M. Grigg, J. K. Weber, R. M. Samstein, V. Makarov, F. Kuo, S. M. Kendall, D. Requena, N. Riaz, B. Greenbaum, J. Carroll, E. Garon, D. M. Hyman, A. Zehir, D. Solit, M. Berger, R. Zhou, N. A. Rizvi, T. A. Chan, Patient HLA class I genotype influences cancer response to checkpoint blockade immunotherapy. *Science*. **359** (2018), pp. 582–587.

74. R. Marty, S. Kaabinejadian, D. Rossell, M. J. Slifker, J. van de Haar, H. B. Engin, N. de Prisco, T. Ideker, W. H. Hildebrand, J. Font-Burgada, H. Carter, MHC-I Genotype Restricts the Oncogenic Mutational Landscape. *Cell*. **171**, 1272–1283.e15 (2017).

75. R. Marty Pyke, W. K. Thompson, R. M. Salem, J. Font-Burgada, M. Zanetti, H. Carter, Evolutionary Pressure against MHC Class II Binding Cancer Mutations. *Cell*. **175**, 1991 (2018).

76. J. M. Vyas, A. G. Van der Veen, H. L. Ploegh, The known unknowns of antigen processing and presentation. *Nat. Rev. Immunol.* **8**, 607–618 (2008).

77. T. Saric, S.-C. Chang, A. Hattori, I. A. York, S. Markant, K. L. Rock, M. Tsujimoto, A. L. Goldberg, An IFN-gamma-induced aminopeptidase in the ER, ERAP1, trims precursors to MHC class I-presented peptides. *Nat. Immunol.* **3**, 1169–1176 (2002).

78. L. D. Hill, D. D. Hilliard, T. P. York, S. Srinivas, J. P. Kusanovic, R. Gomez, M. A. Elovitz, R. Romero, J. F. Strauss 3rd, Fetal ERAP2 variation is associated with preeclampsia in African Americans in a case-control study. *BMC Med. Genet.* **12**, 64 (2011).

79. R. Cagliani, S. Riva, M. Biasin, M. Fumagalli, U. Pozzoli, S. Lo Caputo, F. Mazzotta, L. Piacentini, N. Bresolin, M. Clerici, M. Sironi, Genetic diversity at endoplasmic reticulum aminopeptidases is maintained by balancing selection and is associated with natural resistance to HIV-1 infection. *Hum. Mol. Genet.* **19**, 4705–4714 (2010).

80. S. Liu, D. Cao, Y. Shen, Y. Li, Y. Li, L. Shi, J. Yu, C. Li, X. Zhang, M. Sun, Y. Yao, The ERAP gene is associated with HCV chronic infection in a Chinese Han population. *Hum. Immunol.* **78**, 731–738 (2017).

81. Y. Yao, A. Wiśniewski, Q. Ma, A. Kowal, I. Porębska, K. Pawełczyk, J. Yu, J. Dubis, N. Żuk, Y. Li, L. Shi, P. Kuśnierczyk, Single Nucleotide Polymorphisms of the ERAP1 Gene and Risk of NSCLC: A Comparison of Genetically Distant Populations, Chinese and Caucasian. *Arch. Immunol. Ther. Exp.* . **64**, 117–122 (2016).

82. A. M. Mehta, E. S. Jordanova, T. van Wezel, H.-W. Uh, W. E. Corver, K. M. C. Kwappenberg, W. Verduijn, G. G. Kenter, S. H. van der Burg, G. J. Fleuren, Genetic variation of antigen processing machinery components and association with cervical carcinoma. *Genes Chromosomes Cancer*. **46**, 577–586 (2007).

83. A. M. Mehta, E. S. Jordanova, W. E. Corver, T. van Wezel, H.-W. Uh, G. G. Kenter, G. Jan Fleuren, Single nucleotide polymorphisms in antigen processing machinery component ERAP1 significantly associate with clinical outcome in cervical carcinoma. *Genes Chromosomes Cancer*. **48**, 410–418 (2009).

84. A. M. Mehta, V. M. Spaans, N. B. Mahendra, E. M. Osse, J. N. I. Vet, G. Purwoto, I. G. D. Surya, S. Cornian, A. A. Peters, G. J. Fleuren, E. S. Jordanova, Differences in genetic variation in antigen-processing machinery components and association with cervical carcinoma risk in two Indonesian populations. *Immunogenetics*. **67**, 267–275 (2015).

85. J. Stultz, L. Fong, How to turn up the heat on the cold immune microenvironment of metastatic prostate cancer. *Prostate Cancer Prostatic Dis.* **24**, 697–717 (2021).

86. M. Yarchoan, L. A. Albacker, A. C. Hopkins, M. Montesion, K. Murugesan, T. T. Vithayathil, N. Zaidi, N. S. Azad, D. A. Laheru, G. M. Frampton, E. M. Jaffee, PD-L1 expression and tumor mutational burden are independent biomarkers in most cancers. *JCI Insight*. **4** (2019), doi:10.1172/jci.insight.126908.

87. D. B. Johnson, M. V. Estrada, R. Salgado, V. Sanchez, D. B. Doxie, S. R. Opalenik, A. E. Vilgelm, E. Feld, A. S. Johnson, A. R. Greenplate, M. E. Sanders, C. M. Lovly, D. T. Frederick, M. C. Kelley, A. Richmond, J. M. Irish, Y. Shyr, R. J. Sullivan, I. Puzanov, J. A. Sosman, J. M. Balko, Melanoma-specific MHC-II expression represents a tumour-autonomous phenotype and predicts response to anti-PD-1/PD-L1 therapy. *Nat. Commun.* **7**, 10582 (2016).

88. M. Bilusic, R. A. Madan, J. L. Gulley, Immunotherapy of Prostate Cancer: Facts and Hopes. *Clin. Cancer Res.* **23**, 6764–6770 (2017).

89. P. A. Ott, Z. Hu, D. B. Keskin, S. A. Shukla, J. Sun, D. J. Bozym, W. Zhang, A. Luoma, A.

Giobbie-Hurder, L. Peter, C. Chen, O. Olive, T. A. Carter, S. Li, D. J. Lieb, T. Eisenhaure, E. Gjini, J. Stevens, W. J. Lane, I. Javeri, K. Nellaiappan, A. M. Salazar, H. Daley, M. Seaman, E. I. Buchbinder, C. H. Yoon, M. Harden, N. Lennon, S. Gabriel, S. J. Rodig, D. H. Barouch, J. C. Aster, G. Getz, K. Wucherpfennig, D. Neuberg, J. Ritz, E. S. Lander, E. F. Fritsch, N. Hacohen, C. J. Wu, An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature*. **547**, 217–221 (2017).

90. A. Castro, M. Zanetti, H. Carter, Neoantigen Controversies. *Annu Rev Biomed Data Sci*. **4**, 227–253 (2021).

91. S. E. Brightman, M. S. Naradikian, A. M. Miller, S. P. Schoenberger, Harnessing neoantigen specific CD4 T cells for cancer immunotherapy. *J. Leukoc. Biol.* **107**, 625–633 (2020).

92. E. M. Janssen, E. E. Lemmens, T. Wolfe, U. Christen, M. G. von Herrath, S. P. Schoenberger, CD4+ T cells are required for secondary expansion and memory in CD8+ T lymphocytes. *Nature*. **421**, 852–856 (2003).

93. W. Hemphill, S. Harvey, R. Gumpena, S. Simpson, J. Grayson, T. Smalley, F. Salsbury, T. Hollis, F. Perrino, Small Molecule TREX1 Inhibitors as Novel Immunotherapies. *The FASEB Journal*. **34** (2020), pp. 1–1.

94. P. Chen, Y. Yang, Y. Zhang, S. Jiang, X. Li, J. Wan, Identification of prognostic immune-related genes in the tumor microenvironment of endometrial cancer. *Aging* . **12**, 3371–3387 (2020).

95. M. Espéli, R. Bashford-Rogers, J. M. Sowerby, N. Alouche, L. Wong, A. E. Denton, M. A. Linterman, K. G. C. Smith, FcγRIIb differentially regulates pre-immune and germinal center B cell tolerance in mouse and human. *Nat. Commun.* **10**, 1970 (2019).

96. X. B. Wang, Z. Z. Fan, D. Anton, A. V. Vollenhoven, Z. H. Ni, X. F. Chen, A. K. Lefvert, CTLA4 is expressed on mature dendritic cells derived from human monocytes and influences their maturation and antigen presentation. *BMC Immunol.* **12**, 21 (2011).

97. A. DeRosa, A. Leftin, The Iron Curtain: Macrophages at the Interface of Systemic and Microenvironmental Iron Metabolism and Immune Response in Cancer. *Frontiers in Immunology*. **12** (2021), , doi:10.3389/fimmu.2021.614294.

98. D. Zeng, Z. Ye, J. Wu, R. Zhou, X. Fan, G. Wang, Y. Huang, J. Wu, H. Sun, M. Wang, J. Bin, Y. Liao, N. Li, M. Shi, W. Liao, Macrophage correlates with immunophenotype and predicts anti-PD-L1 response of urothelial cancer. *Theranostics*. **10**, 7002–7014 (2020).

99. C. Kandoth, M. D. McLellan, F. Vandin, K. Ye, B. Niu, C. Lu, M. Xie, Q. Zhang, J. F. McMichael, M. A. Wyczalkowski, M. D. M. Leiserson, C. A. Miller, J. S. Welch, M. J. Walter, M. C. Wendl, T. J. Ley, R. K. Wilson, B. J. Raphael, L. Ding, Mutational landscape and significance across 12 major cancer types. *Nature*. **502**, 333–339 (2013).

100. G. de Velasco, D. Miao, M. H. Voss, A. A. Hakimi, J. J. Hsieh, N. M. Tannir, P. Tamboli, L. J. Appleman, W. K. Rathmell, E. M. Van Allen, T. K. Choueiri, Tumor Mutational Load

and Immune Parameters across Metastatic Renal Cell Carcinoma Risk Groups. *Cancer Immunol Res*. **4**, 820–822 (2016).

101. M. D. Hellmann, T.-E. Ciuleanu, A. Pluzanski, J. S. Lee, G. A. Otterson, C. Audigier-Valette, E. Minenza, H. Linardou, S. Burgers, P. Salman, H. Borghaei, S. S. Ramalingam, J. Brahmer, M. Reck, K. J. O'Byrne, W. J. Geese, G. Green, H. Chang, J. Szustakowski, P. Bhagavatheeswaran, D. Healey, Y. Fu, F. Nathan, L. Paz-Ares, Nivolumab plus Ipilimumab in Lung Cancer with a High Tumor Mutational Burden. *N. Engl. J. Med.* **378**, 2093–2104 (2018).

102. P. C. Tumeh, C. L. Harview, J. H. Yearley, I. P. Shintaku, E. J. M. Taylor, L. Robert, B. Chmielowski, M. Spasic, G. Henry, V. Ciobanu, A. N. West, M. Carmona, C. Kivork, E. Seja, G. Cherry, A. J. Gutierrez, T. R. Grogan, C. Mateus, G. Tomasic, J. A. Glaspy, R. O. Emerson, H. Robins, R. H. Pierce, D. A. Elashoff, C. Robert, A. Ribas, PD-1 blockade induces responses by inhibiting adaptive immune resistance. *Nature*. **515**, 568–571 (2014).

103. M. Ayers, J. Lunceford, M. Nebozhyn, E. Murphy, A. Loboda, D. R. Kaufman, A. Albright, J. D. Cheng, S. P. Kang, V. Shankaran, S. A. Piha-Paul, J. Yearley, T. Y. Seiwert, A. Ribas, T. K. McClanahan, IFN-γ-related mRNA profile predicts clinical response to PD-1 blockade. *J. Clin. Invest.* **127**, 2930–2940 (2017).

104. A. Prat, A. Navarro, L. Paré, N. Reguart, P. Galván, T. Pascual, A. Martínez, P. Nuciforo, L. Comerma, L. Alos, N. Pardo, S. Cedrés, C. Fan, J. S. Parker, L. Gaba, I. Victoria, N. Viñolas, A. Vivancos, A. Arance, E. Felip, Immune-Related Gene Expression Profiling After PD-1 Blockade in Non-Small Cell Lung Carcinoma, Head and Neck Squamous Cell Carcinoma, and Melanoma. *Cancer Res.* **77**, 3540–3550 (2017).

105. A. I. Daud, K. Loo, M. L. Pauli, R. Sanchez-Rodriguez, P. M. Sandoval, K. Taravati, K. Tsai, A. Nosrati, L. Nardo, M. D. Alvarado, A. P. Algazi, M. H. Pampaloni, I. V. Lobach, J. Hwang, R. H. Pierce, I. K. Gratz, M. F. Krummel, M. D. Rosenblum, Tumor immune profiling predicts response to anti-PD-1 therapy in human melanoma. *J. Clin. Invest.* **126**, 3447–3452 (2016).

106. A. C. Huang, M. A. Postow, R. J. Orlowski, R. Mick, B. Bengsch, S. Manne, W. Xu, S. Harmon, J. R. Giles, B. Wenz, M. Adamow, D. Kuk, K. S. Panageas, C. Carrera, P. Wong, F. Quagliarello, B. Wubbenhorst, K. D'Andrea, K. E. Pauken, R. S. Herati, R. P. Staupe, J. M. Schenkel, S. McGettigan, S. Kothari, S. M. George, R. H. Vonderheide, R. K. Amaravadi, G. C. Karakousis, L. M. Schuchter, X. Xu, K. L. Nathanson, J. D. Wolchok, T. C. Gangadhar, E. J. Wherry, T-cell invigoration to tumour burden ratio associated with anti-PD-1 response. *Nature*. **545**, 60–65 (2017).

107. J. Gao, L. Z. Shi, H. Zhao, J. Chen, L. Xiong, Q. He, T. Chen, J. Roszik, C. Bernatchez, S. E. Woodman, P.-L. Chen, P. Hwu, J. P. Allison, A. Futreal, J. A. Wargo, P. Sharma, Loss of IFN-γ Pathway Genes in Tumor Cells as a Mechanism of Resistance to Anti-CTLA-4 Therapy. *Cell*. **167**, 397–404.e9 (2016).

108. N. McGranahan, R. Rosenthal, C. T. Hiley, A. J. Rowan, T. B. K. Watkins, G. A. Wilson, N. J. Birkbak, S. Veeriah, P. Van Loo, J. Herrero, C. Swanton, TRACERx Consortium, Allele-

Specific HLA Loss and Immune Escape in Lung Cancer Evolution. *Cell*. **171**, 1259–1271.e11 (2017).

109. M. Sade-Feldman, Y. J. Jiao, J. H. Chen, M. S. Rooney, M. Barzily-Rokni, J.-P. Eliane, S. L. Bjorgaard, M. R. Hammond, H. Vitzthum, S. M. Blackmon, D. T. Frederick, M. Hazar-Rethinam, B. A. Nadres, E. E. Van Seventer, S. A. Shukla, K. Yizhak, J. P. Ray, D. Rosebrock, D. Livitz, V. Adalsteinsson, G. Getz, L. M. Duncan, B. Li, R. B. Corcoran, D. P. Lawrence, A. Stemmer-Rachamimov, G. M. Boland, D. A. Landau, K. T. Flaherty, R. J. Sullivan, N. Hacohen, Resistance to checkpoint blockade therapy through inactivation of antigen presentation. *Nat. Commun.* **8**, 1136 (2017).

110. J. M. Zaretsky, A. Garcia-Diaz, D. S. Shin, H. Escuin-Ordinas, W. Hugo, S. Hu-Lieskovan, D. Y. Torrejon, G. Abril-Rodriguez, S. Sandoval, L. Barthly, J. Saco, B. Homet Moreno, R. Mezzadra, B. Chmielowski, K. Ruchalski, I. P. Shintaku, P. J. Sanchez, C. Puig-Saus, G. Cherry, E. Seja, X. Kong, J. Pang, B. Berent-Maoz, B. Comin-Anduix, T. G. Graeber, P. C. Tumeh, T. N. M. Schumacher, R. S. Lo, A. Ribas, Mutations Associated with Acquired Resistance to PD-1 Blockade in Melanoma. *N. Engl. J. Med.* **375**, 819–829 (2016).

111. J. C. Maranville, N. J. Cox, Pharmacogenomic variants have larger effect sizes than genetic variants associated with other dichotomous complex traits. *Pharmacogenomics J.* **16**, 388–392 (2016).

112. A. Tin, J. Marten, V. L. Halperin Kuhns, Y. Li, M. Wuttke, H. Kirsten, K. B. Sieber, C. Qiu, M. Gorski, Z. Yu, A. Giri, G. Sveinbjornsson, M. Li, A. Y. Chu, A. Hoppmann, L. J. O'Connor, B. Prins, T. Nutile, D. Noce, M. Akiyama, M. Cocca, S. Ghasemi, P. J. van der Most, K. Horn, Y. Xu, C. Fuchsberger, S. Sedaghat, S. Afaq, N. Amin, J. Ärnlöv, S. J. L. Bakker, N. Bansal, D. Baptista, S. Bergmann, M. L. Biggs, G. Biino, E. Boerwinkle, E. P. Bottinger, T. S. Boutin, M. Brumat, R. Burkhardt, E. Campana, A. Campbell, H. Campbell, R. J. Carroll, E. Catamo, J. C. Chambers, M. Ciullo, M. P. Concas, J. Coresh, T. Corre, D. Cusi, S. C. Felicita, M. H. de Borst, A. De Grandi, R. de Mutsert, A. P. J. de Vries, G. Delgado, A. Demirkan, O. Devuyst, K. Dittrich, K.-U. Eckardt, G. Ehret, K. Endlich, M. K. Evans, R. T. Gansevoort, P. Gasparini, V. Giedraitis, C. Gieger, G. Girotto, M. Gögele, S. D. Gordon, D. F. Gudbjartsson, V. Gudnason, German Chronic Kidney Disease Study, T. Haller, P. Hamet, T. B. Harris, C. Hayward, A. A. Hicks, E. Hofer, H. Holm, W. Huang, N. Hutri-Kähönen, S.-J. Hwang, M. A. Ikram, R. M. Lewis, E. Ingelsson, J. Jakobsdottir, I. Jonsdottir, H. Jonsson, P. K. Joshi, N. S. Josyula, B. Jung, M. Kähönen, Y. Kamatani, M. Kanai, S. M. Kerr, W. Kiess, M. E. Kleber, W. Koenig, J. S. Kooner, A. Körner, P. Kovacs, B. K. Krämer, F. Kronenberg, M. Kubo, B. Kühnel, M. La Bianca, L. A. Lange, B. Lehne, T. Lehtimäki, Lifelines Cohort Study, J. Liu, M. Loeffler, R. J. F. Loos, L.-P. Lyytikäinen, R. Magi, A. Mahajan, N. G. Martin, W. März, D. Mascalzoni, K. Matsuda, C. Meisinger, T. Meitinger, A. Metspalu, Y. Milaneschi, V. A. Million Veteran Program, C. J. O'Donnell, O. D. Wilson, J. M. Gaziano, P. P. Mishra, K. L. Mohlke, N. Mononen, G. W. Montgomery, D. O. Mook-Kanamori, M. Müller-Nurasyid, G. N. Nadkarni, M. A. Nalls, M. Nauck, K. Nikus, B. Ning, I. M. Nolte, R. Noordam, J. R. O'Connell, I. Olafsson, S. Padmanabhan, B. W. J. H. Penninx, T. Perls, A. Peters, M. Pirastu, N. Pirastu, G. Pistis, O. Polasek, B. Ponte, D. J. Porteous, T. Poulain, M. H. Preuss, T. J. Rabelink, L. M. Raffield, O. T. Raitakari, R. Rettig, M. Rheinberger, K. M. Rice, F. Rizzi, A. Robino, I. Rudan, A. Krajcoviechova, R. Cifkova, R.

Rueedi, D. Ruggiero, K. A. Ryan, Y. Saba, E. Salvi, H. Schmidt, R. Schmidt, C. M. Shaffer, A. V. Smith, B. H. Smith, C. N. Spracklen, K. Strauch, M. Stumvoll, P. Sulem, S. M. Tajuddin, A. Teren, J. Thiery, C. H. L. Thio, U. Thorsteinsdottir, D. Toniolo, A. Tönjes, J. Tremblay, A. G. Uitterlinden, S. Vaccargiu, P. van der Harst, C. M. van Duijn, N. Verweij, U. Völker, P. Vollenweider, G. Waeber, M. Waldenberger, J. B. Whitfield, S. H. Wild, J. F. Wilson, Q. Yang, W. Zhang, A. B. Zonderman, M. Bochud, J. G. Wilson, S. A. Pendergrass, K. Ho, A. Parsa, P. P. Pramstaller, B. M. Psaty, C. A. Böger, H. Snieder, A. S. Butterworth, Y. Okada, T. L. Edwards, K. Stefansson, K. Susztak, M. Scholz, I. M. Heid, A. M. Hung, A. Teumer, C. Pattaro, O. M. Woodward, V. Vitart, A. Köttgen, Target genes, variants, tissues and transcriptional pathways influencing human serum urate levels. *Nat. Genet.* **51**, 1459–1474 (2019).

113. H. Carter, R. Marty, M. Hofree, A. M. Gross, J. Jensen, K. M. Fisch, X. Wu, C. DeBoever, E. L. Van Nostrand, Y. Song, E. Wheeler, J. F. Kreisberg, S. M. Lippman, G. W. Yeo, J. S. Gutkind, T. Ideker, Interaction Landscape of Inherited Polymorphisms with Somatic Events in Cancer. *Cancer Discov.* **7**, 410–423 (2017).

114. J. M. Korn, F. G. Kuruvilla, S. A. McCarroll, A. Wysoker, J. Nemesh, S. Cawley, E. Hubbell, J. Veitch, P. J. Collins, K. Darvishi, C. Lee, M. M. Nizzari, S. B. Gabriel, S. Purcell, M. J. Daly, D. Altshuler, Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat. Genet.* **40**, 1253–1260 (2008).

115. S. Das, L. Forer, S. Schönherr, C. Sidore, A. E. Locke, A. Kwong, S. I. Vrieze, E. Y. Chew, S. Levy, M. McGue, D. Schlessinger, D. Stambolian, P.-R. Loh, W. G. Iacono, A. Swaroop, L. J. Scott, F. Cucca, F. Kronenberg, M. Boehnke, G. R. Abecasis, C. Fuchsberger, Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).

116. R. M. Kuhn, D. Haussler, W. J. Kent, The UCSC genome browser and associated tools. *Brief. Bioinform.* **14**, 144–161 (2013).

117. C. Bellenguez, A. Strange, C. Freeman, Wellcome Trust Case Control Consortium, P. Donnelly, C. C. A. Spencer, A robust clustering algorithm for identifying problematic samples in genome-wide association studies. *Bioinformatics*. **28**, 134–135 (2012).

118. A. M. Newman, C. B. Steen, C. L. Liu, A. J. Gentles, A. A. Chaudhuri, F. Scherer, M. S. Khodadoust, M. S. Esfahani, B. A. Luca, D. Steiner, M. Diehn, A. A. Alizadeh, Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat. Biotechnol.* **37**, 773–782 (2019).

119. J. Yang, R. J. F. Loos, J. E. Powell, S. E. Medland, E. K. Speliotes, D. I. Chasman, L. M. Rose, G. Thorleifsson, V. Steinthorsdottir, R. Mägi, L. Waite, A. V. Smith, L. M. Yerges-Armstrong, K. L. Monda, D. Hadley, A. Mahajan, G. Li, K. Kapur, V. Vitart, J. E. Huffman, S. R. Wang, C. Palmer, T. Esko, K. Fischer, J. H. Zhao, A. Demirkan, A. Isaacs, M. F. Feitosa, J. Luan, N. L. Heard-Costa, C. White, A. U. Jackson, M. Preuss, A. Ziegler, J. Eriksson, Z. Kutalik, F. Frau, I. M. Nolte, J. V. Van Vliet-Ostaptchouk, J.-J. Hottenga, K. B. Jacobs, N. Verweij, A. Goel, C. Medina-Gomez, K. Estrada, J. L. Bragg-Gresham, S. Sanna, C. Sidore, J. Tyrer, A. Teumer, I. Prokopenko, M. Mangino, C. M. Lindgren, T. L. Assimes, A. R.

Shuldiner, J. Hui, J. P. Beilby, W. L. McArdle, P. Hall, T. Haritunians, L. Zgaga, I. Kolcic, O. Polasek, T. Zemunik, B. A. Oostra, M. Juhani Junttila, H. Grönberg, S. Schreiber, A. Peters, A. A. Hicks, J. Stephens, N. S. Foad, J. Laitinen, A. Pouta, M. Kaakinen, G. Willemsen, J. M. Vink, S. H. Wild, G. Navis, F. W. Asselbergs, G. Homuth, U. John, C. Iribarren, T. Harris, L. Launer, V. Gudnason, J. R. O'Connell, E. Boerwinkle, G. Cadby, L. J. Palmer, A. L. James, A. W. Musk, E. Ingelsson, B. M. Psaty, J. S. Beckmann, G. Waeber, P. Vollenweider, C. Hayward, A. F. Wright, I. Rudan, L. C. Groop, A. Metspalu, K.-T. Khaw, C. M. van Duijn, I. B. Borecki, M. A. Province, N. J. Wareham, J.-C. Tardif, H. V. Huikuri, L. Adrienne Cupples, L. D. Atwood, C. S. Fox, M. Boehnke, F. S. Collins, K. L. Mohlke, J. Erdmann, H. Schunkert, C. Hengstenberg, K. Stark, M. Lorentzon, C. Ohlsson, D. Cusi, J. A. Staessen, M. M. Van der Klauw, P. P. Pramstaller, S. Kathiresan, J. D. Jolley, S. Ripatti, M.-R. Jarvelin, E. J. C. de Geus, D. I. Boomsma, B. Penninx, J. F. Wilson, H. Campbell, S. J. Chanock, P. van der Harst, A. Hamsten, H. Watkins, A. Hofman, J. C. Witteman, M. Carola Zillikens, A. G. Uitterlinden, F. Rivadeneira, L. A. Kiemeney, S. H. Vermeulen, G. R. Abecasis, D. Schlessinger, S. Schipf, M. Stumvoll, A. Tönjes, T. D. Spector, K. E. North, G. Lettre, M. I. McCarthy, S. I. Berndt, A. C. Heath, P. A. F. Madden, D. R. Nyholt, G. W. Montgomery, N. G. Martin, B. McKnight, D. P. Strachan, W. G. Hill, H. Snieder, P. M. Ridker, U. Thorsteinsdottir, K. Stefansson, T. M. Frayling, J. N. Hirschhorn, M. E. Goddard, P. M. Visscher, FTO genotype is associated with phenotypic variability of body mass index. *Nature*. **490**, 267–272 (2012).

120. J. Yang, S. H. Lee, M. E. Goddard, P. M. Visscher, GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).

121. J. Yang, B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, M. E. Goddard, P. M. Visscher, Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).

122. S. Seabold, J. Perktold, Statsmodels: Econometric and Statistical Modeling with Python. *Proceedings of the 9th Python in Science Conference* (2010), , doi:10.25080/majora-92bf1922-011.

123. Y.-B. Zhang, J. Hu, J. Zhang, X. Zhou, X. Li, C. Gu, T. Liu, Y. Xie, J. Liu, M. Gu, P. Wang, T. Wu, J. Qian, Y. Wang, X. Dong, J. Yu, Q. Zhang, Genome-wide association study identifies multiple susceptibility loci for craniofacial microsomia. *Nat. Commun.* **7**, 10605 (2016).

124. V. R. C. Aguiar, J. César, O. Delaneau, E. T. Dermitzakis, D. Meyer, Expression estimation and eQTL mapping for HLA genes with a personalized pipeline. *PLoS Genet.* **15**, e1008091 (2019).

125. A. Frankish, M. Diekhans, A.-M. Ferreira, R. Johnson, I. Jungreis, J. Loveland, J. M. Mudge, C. Sisu, J. Wright, J. Armstrong, I. Barnes, A. Berry, A. Bignell, S. Carbonell Sala, J. Chrast, F. Cunningham, T. Di Domenico, S. Donaldson, I. T. Fiddes, C. García Girón, J. M. Gonzalez, T. Grego, M. Hardy, T. Hourlier, T. Hunt, O. G. Izuogu, J. Lagarde, F. J. Martin, L. Martínez, S. Mohanan, P. Muir, F. C. P. Navarro, A. Parker, B. Pei, F. Pozo, M. Ruffier,

B. M. Schmitt, E. Stapleton, M.-M. Suner, I. Sycheva, B. Uszczynska-Ratajczak, J. Xu, A. Yates, D. Zerbino, Y. Zhang, B. Aken, J. S. Choudhary, M. Gerstein, R. Guigó, T. J. P. Hubbard, M. Kellis, B. Paten, A. Reymond, M. L. Tress, P. Flicek, GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).

126. W. McLaren, L. Gil, S. E. Hunt, H. S. Riat, G. R. S. Ritchie, A. Thormann, P. Flicek, F. Cunningham, The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).

127. D. Aran, Z. Hu, A. J. Butte, xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol.* **18**, 220 (2017).

128. B. J. Schmiedel, D. Singh, A. Madrigal, A. G. Valdovino-Gonzalez, B. M. White, J. Zapardiel-Gonzalo, B. Ha, G. Altay, J. A. Greenbaum, G. McVicker, G. Seumois, A. Rao, M. Kronenberg, B. Peters, P. Vijayanand, Impact of Genetic Polymorphisms on Human Immune Cell Gene Expression. *Cell*. **175**, 1701–1715.e16 (2018).

129. G. Abraham, Y. Qiu, M. Inouye, FlashPCA2: principal component analysis of Biobank-scale genotype datasets. *Bioinformatics*. **33**, 2776–2778 (2017).

130. M. A. Hall, J. Wallace, A. Lucas, D. Kim, A. O. Basile, S. S. Verma, C. A. McCarty, M. H. Brilliant, P. L. Peissig, T. E. Kitchner, A. Verma, S. A. Pendergrass, S. M. Dudek, J. H. Moore, M. D. Ritchie, PLATO software provides analytic framework for investigating complexity beyond genome-wide association studies. *Nat. Commun.* **8**, 1167 (2017).

131. J. Liu, T. Lichtenberg, K. A. Hoadley, L. M. Poisson, A. J. Lazar, A. D. Cherniack, A. J. Kovatich, C. C. Benz, D. A. Levine, A. V. Lee, L. Omberg, D. M. Wolf, C. D. Shriver, V. Thorsson, Cancer Genome Atlas Research Network, H. Hu, An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell*. **173**, 400–416.e11 (2018).

132. D. Miao, C. A. Margolis, W. Gao, M. H. Voss, W. Li, D. J. Martini, C. Norton, D. Bossé, S. M. Wankowicz, D. Cullen, C. Horak, M. Wind-Rotolo, A. Tracy, M. Giannakis, F. S. Hodi, C. G. Drake, M. W. Ball, M. E. Allaf, A. Snyder, M. D. Hellmann, T. Ho, R. J. Motzer, S. Signoretti, W. G. Kaelin Jr, T. K. Choueiri, E. M. Van Allen, Genomic correlates of response to immune checkpoint therapies in clear cell renal cell carcinoma. *Science*. **359**, 801–806 (2018).

133. H. Li, R. Durbin, Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*. **26**, 589–595 (2010).

134. H. Li, A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*. **27**, 2987–2993 (2011).

135. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. **25**, 2078–2079 (2009).

136. G. A. Auwera, M. O. Carneiro, C. Hartl, R. Poplin, G. del Angel, A. Levy-Moonshine, T. Jordan, K. Shakir, D. Roazen, J. Thibault, E. Banks, K. V. Garimella, D. Altshuler, S. Gabriel, M. A. DePristo, From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Current Protocols in Bioinformatics*. **43** (2013), , doi:10.1002/0471250953.bi1110s43.

137. M. A. DePristo, E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M. Kernytsky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler, M. J. Daly, A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).

138. A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, M. A. DePristo, The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).

139. R. Poplin, P.-C. Chang, D. Alexander, S. Schwartz, T. Colthurst, A. Ku, D. Newburger, J. Dijamco, N. Nguyen, P. T. Afshar, S. S. Gross, L. Dorfman, C. Y. McLean, M. A. DePristo, A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **36**, 983–987 (2018).

140. T. Yun, H. Li, P.-C. Chang, M. F. Lin, A. Carroll, C. Y. McLean, Accurate, scalable cohort variant calls using DeepVariant and GLnexus, , doi:10.1101/2020.02.10.942086.

141. L. Seymour, J. Bogaerts, A. Perrone, R. Ford, L. H. Schwartz, S. Mandrekar, N. U. Lin, S. Litière, J. Dancey, A. Chen, F. S. Hodi, P. Therasse, O. S. Hoekstra, L. K. Shankar, J. D. Wolchok, M. Ballinger, C. Caramella, E. G. E. de Vries, RECIST working group, iRECIST: guidelines for response criteria for use in trials testing immunotherapeutics. *Lancet Oncol.* **18**, e143–e152 (2017).

142. J. A. Edwards, R. A. Edwards, Fastq-pair: efficient synchronization of paired-end fastq files, , doi:10.1101/552885.

143. A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, T. R. Gingeras, STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. **29**, 15–21 (2013).

144. B. Li, C. N. Dewey, RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*. **12** (2011), , doi:10.1186/1471-2105-12-323.

CHAPTER 2: TIME-SNPs in Pediatric Cancer

2.1.1 Abstract

Pediatric tumors are the second leading cause of death in children. Furthermore, pediatric cancer survivors experience a sequela of long-term effects related to their cancer and treatment that are debilitating. Immunotherapy is a promising alternative for pediatric patients; however, immune checkpoint blockade clinical trials in children show limited benefit and a better understanding of the pediatric tumor immune microenvironment (TIME) is needed. Since pediatric immunotherapy trial data is limited, we evaluated known determinants of the tumor immune microenvironment and compared the relevance in adult and pediatric tumors. To address this, we started by doing a comprehensive analysis of over 2000 pediatric patients and compared genetic determinants of antigen presentation in pediatric and adult patients. One of the main antigen presentation machineries is HLA and we found generally HLA genotype allele frequencies were consistent between pediatric and adult tumors. We conducted analyses to identify enriched or depleted HLA alleles in pediatric and adult cancers compared to non-cancer controls in the UK Biobank. Identified HLA alleles enriched or depleted in pediatric and adult cancer were shared. Several pediatric cancer-associated HLA alleles were associated with age of diagnosis and a minority of HLA alleles significantly influenced immune infiltration in an age-dependent manner. Adult antigen presentation TIME-SNPs were significantly associated with antigen presentation in pediatric tumors as well. These results suggest germline determinants of antigen presentation are shared between adults and pediatric patients, suggesting that host anti-tumor immunity is playing a role in pediatric cancer risk as well.

2.1.2 Introduction

Pediatric tumors are the second leading cause of death in children after accidents(*1*). The most common types of cancer include leukemias, central nervous system (CNS) tumors, lymphomas, neuroblastomas, kidney tumors and malignant bone tumors. An improvement in pediatric cancer survival has been observed between 1994 and 2014; however, survival rates for some pediatric cancer types remain low. For example, pediatric brain tumors have replaced leukemias as the leading cause of pediatric cancer death(*2–4*). Furthermore, pediatric cancer survivors still face challenges related to cancer treatment such as cognitive and developmental deficits, endocrine disorders and an increased susceptibility to secondary cancers(*5*). A better understanding of pediatric cancers is needed to provide better care to pediatric cancer patients and address long-term complications in pediatric cancer survivors.

Pediatric tumors are different from adult tumors. Adult tumors are relatively more frequent compared to pediatric tumors. Approximately 10,600 cases of cancer are diagnosed in children under age 15 compared to more than 1.7 million in adults in the US every year(*6*). Types of tumors are also distinct, with 50% and 26% of pediatric tumors being leukemias or lymphomas and central nervous tumors, respectively, whereas in adults, the corresponding rates are 10% and 1.5%. Pediatric tumors are considered more aggressive and progress faster, with 80% of pediatric tumors detected after they have already metastasized.

Genetically, pediatric tumors tend to have fewer mutations and different types of mutations. The lower number of somatic mutations in pediatric cancer has been attributed to the embryonic origin of these cancers, developmental gene dysregulation and fewer environmental exposures, although pediatric tumors with alterations in DNA damage repair pathway can have higher rates of mutations(*7*, *8*). In a pan-cancer pediatric genetic analysis, 142 pediatric cancer

driver genes were identified but only half overlapped with adult cancer driver genes(*9*). Despite comprehensive genetic analysis(*10*, *11*), the causes of pediatric tumors usually are unknown.

Pediatric tumors do tend to respond better to therapeutic options such as chemotherapy and radiation compared to adults; however, these therapeutics can have long-term consequences. Pediatric cancer survivors, for example, have higher risk of congestive heart failure, coronary artery disease, secondary cancers, anxiety, and depression. Immunotherapy can serve as a promising avenue to avoid these long-term side effects of cancer treatment; however, most pediatric clinical trials using immunotherapy have not indicated a significant benefit(*12–15*). Better understanding the extent and nature of immune involvement in pediatric tumors may reveal why therapies that work for adults fail or children and may reveal better strategies to target pediatric tumors with the immune system. Furthermore, most immunotherapy biomarkers have been identified in adult cohorts and better biomarkers for immunotherapy are needed in pediatric patients, especially considering the large differences between pediatric and adult cancer. Unfortunately, the number of pediatric immunotherapy trials are limited, making biomarker discovery difficult. To address this, we used methods for characterizing the adult TIME and applied them to pediatric tumors to determine their relevance. We specifically focused on applying methods for assessing antigen presentation. We characterized HLA genotype and antigen presentation variants in pediatric tumors, demonstrating that many of these characteristics of the adult TIME are shared with pediatric tumors.

2.1.3 Results

2.1.3.1 Collection of Pediatric Multi-Omics Data

Pediatric samples were acquired from St. Jude Pediatric Cancer Genome Project, Kids' First Children's Brain Tumor Atlas, TARGET, and International Cancer Genome Consortium and harmonized, resulting in whole genome germline SNPs, exonic somatic SNPs and RNA gene counts for analysis. After processing, we had germline information for 2288 pediatric blood tumors and 2664 pediatric solid tumors, of which 972 were pediatric brain tumors. For immune microenvironment analysis, we focused on pediatric solid and brain tumors **(Figure 2.1, S2.1)**. PCA analysis on intersecting variants was conducted with 1000 Genomes individuals to infer genetic ancestry **(Figure 2.2)**. 2483, 412, 104 and 50 individuals were of European, African, East Asian and South Asian ancestry based on k-means analysis, respectively **(Figure 2.2)**.

To ensure the quality of germline variants calls, we compared calls with previous analysis done by Zhang *et al.* for pathogenic variants in cancer predisposition genes. Of the 52 pathogenic TP53 variants identified by Zhang *et al.*, we found 42 of the TP53 variants. The missing 10 variants were due to data being unavailable for certain samples.

2.1.3.2 HLA Profiling of Pediatric Tumors Identified Alleles Specifically Associated with Pediatric Cancer and Age of Diagnosis

Antigen presentation is an important immune process mediating anti-tumor activities targeted against mutant proteins harbored by tumor cells. The major histocompatibility complex (MHC) is a critical component of the antigen presentation machinery that displays peptides on the cell surface for immune surveillance by T cells. The human leukocyte antigone (HLA) genes encoding the peptide binding groove of both MHC Class I and II molecules are located on chromosome 6. Previous studies have demonstrated MHC Class I and II genotypes can determine whether immune surveillance can act against somatic driver mutations(*16*, *17*). This process is

thought to be a critical determinant of response to immunotherapy in the adult setting. Among adult tumors with high tumor mutation burden, those with at least one driver mutation effectively presented by MHC-I were more likely to benefit from immunotherapy(*18*). However in the pediatric setting where tumor mutation burden is low, availability of neoantigens is more limited. Furthermore, the pediatric adaptive immune system is still developing and thus more variation is observed in immune responses compared to adults(*19*, *20*). To assess the role of antigen presentation in pediatric patients, we first conducted comprehensive HLA profiling of pediatric patients.

We called HLA types for pediatric cancer patients and evaluated the frequency of HLA-A, HLA-B, HLA-C genotypes in our pediatric cohort and compared them with frequencies in the adult cancer population. Generally, allele frequencies were highly concordant in both populations **(Figure 2.3)**. Based on Fisher's exact analysis, HLA-A23:01, HLA-B48:01, HLA-B56:01 and HLA-C18:01 were overrepresented in pediatric patients compared to the adult TCGA cohort, while HLA-A01:03, HLA-B27:02, HLA-C12:03 were depleted in pediatric patients compared to adult patients. Among these, HLA-B56:01 is a known autoimmune allele associated with ankylosing spondylitis and psoriasis (*21*). Enrichment of HLA-B56:01 may suggest that dysregulation of the immune system can influence pediatric cancer risk.

We next wanted to determine if there was a difference in HLA alleles enriched in adult cancer compared to pediatric cancer. We looked at enriched or depleted HLA alleles in the pediatric St. Jude PCPG cohort and TCGA compared to noncancer controls in the UK Biobank and observed high overlap **(Figure 2.3)**. Several of these HLA alleles have been implicated in cancer risk based on the Vanderbilt HLA PheWAS catalog. Only 5 HLA alleles (HLA-A30:02, HLA-B14:02, HLA-B57:02, HLA-B39:02, HLA-C07:04) were uniquely enriched in pediatric

cancer patients. Odds ratios of identified HLA alleles were consistent between adult and pediatric cancer patients (**Figure 2.3).** We identified enriched and depleted HLA alleles by cancer type and found a large number of HLA alleles enriched or depleted in neuroblastoma**(Figure 2.3).** Neuroblastoma did have the highest number of samples and 2 of the alleles overlapped with cancer type-specific alleles identified for high grade glioma and medulloblastoma (**Figure 2.3).** Three of the neuroblastoma-associated alleles were associated with cancer in the PheWAS catalog **(Figure 2.3)**.

Since the immune system is only fully developed at age 12, we hypothesized that HLA genotypes may be significantly associated with age. Indeed, ~68% of HLA alleles were associated with age of diagnosis, controlling for top 10 principal components from genotype PCA analysis **(Figure 2.3).** We next identified HLA alleles which were associated with immune infiltration, suggesting they may play a role in shaping the tumor immune microenvironment. After controlling for multiple hypothesis testing, 7.7% of alleles were associated with immune infiltration **(Figure 2.3).** Since many of the HLA alleles were associated with age of diagnosis, we wanted to determine if the effect of HLA alleles on immune infiltration was influenced by age. Indeed, ~4.5% of alleles had significant interactions with age of diagnosis (**Figure 2.3).** One example is HLA-A33:03 which was significantly associated with monocyte infiltration in pediatric cancer patients > 12 years, but not those < 12 years of age **(Figure 2.3)**. These results suggest that HLA genotype plays a significant role in pediatric cancer and the immune microenvironment and this effect may be influenced by the TIME. Since early development requires immune suppression potentially driven by Treg cells, antigen presentation may have more of an effect in older pediatric cancer patients.

We next identified germline and somatic variants in the St. Jude PCPG cohort that affected antigen presentation. No documented ClinVar pathogenic germline variants were associated with

HLA genes; however, 43 patients had somatic HLA mutations resulting in protein sequence changes. HLA mutation was not significantly associated with age of diagnosis overall; however, in high grade glioma, HLA mutation was associated with later age of diagnosis **(Figure 2.4)**. HGG patients had the following mutations: HLA-A mutation - p.G231S, HLA-B mutation - p.H27R, p.A348V, p.W298R, HLA-C mutation - p.T282M. HLA mutation can serve as a mechanism of immune evasion through reduced antigen presentation and remodeling of the antigen landscape. The presence of HLA mutations in pediatric tumors indicate that immune evasion may play a role in pediatric tumors as well.

2.1.3.3 Adult Tumor Immune Microenvironment (TIME) Variants also Modify Pediatric TIME

To determine the effects of other immune microenvironment modifiers in pediatric cancer patients, we ran association studies with adult TIME-SNPs identified from Aim 1 and SNP-heritable IP components. We found several of these variants were also significantly associated with IP components in pediatric cancer (**Figure 2.5**). We could only assess 791 of the 1084 variants due to the smaller cohort size we had for pediatric cancers. Of the 791 variants, 76 were significantly associated with IP components (FDR < .05), *ENTPD1, ERAP1, ERAP2, DHFR, DCTN5, CTSW, BTN3A2, MICA, DCK* and MHC Class II genes. Indeed, these genes are enriched for antigen processing and presentation functions (GO Fold enrichment =98.93, FDR<2.38E-11).

Given the number of adult TIME-SNPs associated with antigen presentation which were significant in pediatric cancer patients, we sought to determine their role in pediatric patients. From our adult germline analysis, we discovered an antigen presentation PRS associated with melanoma risk. To determine the relevance of this PRS in a pediatric population, we constructed the antigen

presentation PRS in the pediatric population. We found that the distribution of scores were similar to noncancer adults in the UK Biobank and noncancer pediatric patients in SPARK cohort(*22*). We did find the PRS was significantly lower in medulloblastoma patients compared to the UK Biobank and SPARK autism cohort **(Figure 2.5)**. PRS was not associated with age of diagnosis or immune infiltration. With a more relaxed FDR (FDR < 0.2), PRS was associated with activated mast cells and regulatory T cells; however, larger sample sizes are likely needed to validate these findings.

2.1.3.4 DNA-damage repair (DDR) Variants Underlying Pediatric TIME

Pediatric patients with mismatch repair and polymerase proofreading deficiencies have higher mutational loads(*23*). In adult cancers, TMB is frequently used as a genomic predictor of immune checkpoint blockade response. Analysis of International Replication Repair Deficiency Consortium pediatric patients treated with anti-PD-1 revealed that hypermutant pediatric cancer patients were more likely to be responders(*24*). To assess the role of germline MMR pathogenic variants in the pediatric TIME, we first calculated burden of 16 homologous repair (HR) and nucleotide excision repair (NER) genes(*25*). We identified 70 patients with a HR/NER pathogenic germline variants (**Figure 2.6).** Genes with largest number of pathogenic variants include *RAD45L* (19), *BRCA2* (15), and *ERCC2* (11). Germline HR/NER pathogenic variants were not associated with mutational burden in all pediatric patients but was associated with increased mutation burden in medulloblastoma patients **(Figure 2.6)**.

We next evaluated association of mutational burden in an age-dependent manner. We observed higher mutational load in older pediatric patients (>12 years) compared to younger pediatric patients (<12 years). We then ran associations with CIBERSORTx immune infiltrate

estimate and DDR variant status in older and younger pediatric patients and only observed significant association in older pediatric patients. For example, effects of DDR variant on M2 macrophage infiltration were significantly different in older and younger pediatric patients **(Figure 2.6).**

2.1.3.5 Single-Cell Profiling Cell Infiltrate Estimates are Distinct from Default Methods

LM22 default matrix is composed of 547 genes to distinguish 22 mature human heamtopoeitic populations from peripheral blood (*26*). Although LM22 signature matrix has been validated, it may miss distinct cells only present in solid tumors, such as microglia, and may not accurately reflect the tumor immune microenvironment. To address this, we constructed a custom signature matrix from scRNA-seq available from Gojo et al (*27*) and estimated cell infiltrates. We compared immune cell infiltrate estimates with LM22 immune cell infiltrate estimates in pediatric ependymoma samples and found low correlation ($\rho < 0.5$) **(Figure 2.7)**. We next sought to understand how these immune cell infiltrate estimates could affect downstream analyses. M2 Macrophages and microglia had some correlation in estimate values, so we ran germline associations with M2 Macrophages and microglia and compared suggestive association ($p < 1e-05$). No overlap was observed suggesting that signature matrix selection can critically influence results

2.1.4 Discussion

Pediatric tumors are rare and distinct from adult tumors. Although survival rates have improved for pediatric tumors, patients suffer from long-term effects of treatment and better options for care are needed. Immunotherapy is a promising alternative; however, pediatric

immunotherapy clinical trials have not been successful. In this study, we wanted to evaluate whether germline factors associated with anti-tumor immunity and the tumor immune microenvironment in adults were also present in pediatric disease.

We started by profiling the HLA region in more than 4000 pediatric tumors. We found a number of HLA alleles had different frequencies in cancer populations relative to expectation for matched populations in both pediatric and adult disease. We also found several HLA alleles were associated with pediatric age at cancer diagnosis even after controlling for tumor type and influenced immune cell infiltration in an age-dependent manner. These findings may correlate with early stages of immunosuppression in pediatric patients that changes after immune system development.

We also evaluated adult TIME-SNP associations in the pediatric cohort and found adult TIME-SNPs associated with antigen presentation influenced antigen presentation in pediatric tumors also. These results suggest that germline variants underlying the adult and pediatric TIME related to antigen presentation were consistent. Although unique cancer risk alleles for pediatric cancer were identified, the majority of HLA alleles and TIME-SNPs had similar behavior in a pediatric cancer setting.

Germline studies in pediatric cancer are limited, because pediatric cancer is relatively rare and finding large sample sizes powered for association studies is difficult. Here, we combined results from more than 4 genetically profiled pediatric cancer cohorts and conducted comprehensive TIME profiling. We present a comprehensive HLA allele profiling studies in pediatric cancer and find that HLA genotypes can significantly influence age of diagnosis in pediatric patients. We also find an enrichment of autoimmune allele HLA-B56:01 in pediatric cancer compared to adult cancer, indicating the dysregulation of the immune system can influence

pediatric cancer risk. We only used one tool for HLA profiling; however, consensus calls from multiple HLA genotype callers is a future direction to confirm results.

Immunoprofiling studies of the pediatric TIME have found high macrophage infiltration, low dendritic cell infiltration and low levels of immune checkpoint expression(*26–28*). Through our adult germline TIME analysis, we found a large number of antigen presentation-related TIME-SNPs to be associated with cancer risk and immunotherapy response in adults. In our study, we found a subset of TIME-SNPs which also influenced antigen presentation machinery expression in the pediatric setting. Antigen presentation is a crucial immune process in humans and can shape tumor behavior. The similarity in significant association with antigen presentation in pediatric and adult cancer patients suggest that more investigation into antigen presentation may be fruitful in pediatric cancer patients.

Future directions include exploring other immune microenvironment characteristics, such as TMB, neoantigen burden and immune checkpoint expression, in our pediatric cancer cohort. Given the number of HLA alleles correlated with age of diagnosis, incorporating age as a variable into analyses will be important. Although we find several adult TIME-SNPs were associated with pediatric TIME, we note that our pediatric cohort is still small, limiting the number of analyses that can be conducted. Hopefully, with processing of more pediatric cancer cohort, more associations can be captured and linked to the TIME.

2.1.5 Materials and Methods

2.1.5.1 Pediatric Subject Details

St. Jude Pediatric Cancer Genome Project (PCPG)(*26*, *27*), Kids First Children's Brain Tumor Network (CBTN), Target and International Cancer Genome Consortium (ICGC) had tumor

and matched normal samples for over 6000 individuals (age < 20 years). St. Jude PCPG information was accessed from DNAnexus (project-Fk1P9x09ZgZJbb0q38XffZJB), ICGC information for PBCA-DE and PEME-DE were accessed through ICGC data portal (https://dcc.icgc.org/) using score client, CBTN information was accessed through data portal (https://portal.kidsfirstdrc.org/dashboard) using Seven bridges and Target information was accessed through Genomic Data Commons (GDC) archive using gdc client. Individuals were consented for general research use and no attempts were made to reidentify or contact subjects. Both females and males were included, and sex and individual age were included as covariates. More detailed subject information can be found in Supplementary Figure 1.

2.1.5.2 Pediatric Genotypes

For PCPG individuals with WGS, normal gVCFs were downloaded, merged with GLnexus[63] and converted to PLINK format for analysis of genotypes. For PCPG individuals with only WXS, normal gVCFs were downloaded, merged with GLnexus[63] and converted to PLINK format for analysis of genotypes. Only biallelic variants chromosomes 1-22, X and Y were kept. SNPs with call rate <90% and minor allele frequency (MAF) <1% were removed. Individuals with genotype coverage <90% were removed. After applying these filters, the remaining 565,487 SNPs were input to the secure TOPMED server. SNPs were imputed with Minimac3/Minimac4 and TOPMED reference with Eaglev2.3 phasing. For CBTN and ICGC individuals, normal WGS gVCFs were downloaded, merged with GLNexus and converted to PLINK format for analysis like PCPG processing. For Target individuals, normal (non-tumor) level 2 genotype calls generated from Affymetrix SNP6.0 array intensities using BIRDSUITE software were retrieved from GDC Legacy Portal (accession date: 02/28/2022). We discarded copy number intensity calls or

Affymetrix SNP probs with no associated SNP rsid. In these files, each of the remaining 932,824 SNPs was annotated with an allele count (0=AA, 1=AB, 2=BB, -1=missing) and confidence score between 0 and 1. Genotypes with a score larger than 0.1 (error rate > 10%) were set to missing and data were reformatted for PLINK (*28*). Allele counts were converted to alleles using the definitions in metadata distributed with Affymetrix SNP 6.0 Array Documentation and negative strand genotypes were flipped to the positive strand using PLINK (PLINK, RRID:SCR_001757).(*28*) SNPs with call rate <90% and minor allele frequency (MAF) <1% were removed. Individuals with genotype coverage <90% were removed. After applying these filters, the remaining 844,265 SNPs were input to the secure TOPMED server.

After imputation or gVCF compilation, the remaining SNP counts were: 52,622,417 for PCPG WGS (reference: hg38), 63,847,762 for PCPG WGS (reference: GRCh38), 53,198,657 for PCPG WXS (reference: GRCh38), 60,592,831 for CBTN, 21,882,353 for ICGC PBCA-DE, 22,258,387 for ICGC PEME-DE, 292,136,462 for Target.

2.1.5.3 Pediatric Population Stratification

1000 Genomes genotypes for 2,504 individuals were accessed from https://www.cog-genomics.org/plink/2.0/resources#1kg_phase (accession date: 02/27/21) and lifted to hg38 using the liftOver utility. Intersecting genotypes between 1000 Genomes and pediatric genotypes were extracted and used for pairwise identity-by-state (IBS) analysis in PLINK. K-means clustering trained on 1000 Genomes individuals separated into the following groups: EUR, AFR, AMR, EAS and SAS. This trained model was used to predict groups in TCGA. Principal components from IBS analysis were used as covariates for germline analysis.

2.1.5.4 Pediatric TMB

For TMB, somatic VCFs were acquired for all pediatric patients. The number of somatic mutations resulting in protein sequence changes in exonic regions was quantified and used for analysis.

2.1.5.5 Pediatric HLA typing

HLA genotyping and mutation identification was performed for genes *HLA-A*, *HLA-B* and *HLA-C* using PolySolver (with default parameters)(*29*).

HLA enrichment and depletion was calculated with Fisher's exact test. For adult and pediatric cancer comparison, TCGA and pediatric HLA alleles were compared. For pediatric and adult cancer allele identification, Fisher's exact test between TCGA and pediatric cancer HLA calls with UK Biobank non-cancer calls was conducted. For association with age, linearregression with age of diagnosis and HLA type and top 10 principal components was conducted. For associations with immune infiltration, pediatric RNA was processed using RSEM 1.3.1. Expected counts were transformed to TPM and used as input for CIBERSORTx(*30*). LM22 signature matrix was used for deconvolution and association with immune infiltration. Benjamini-Hochberg FDR calculation was used for multiple hypothesis testing.

2.1.6 Figures

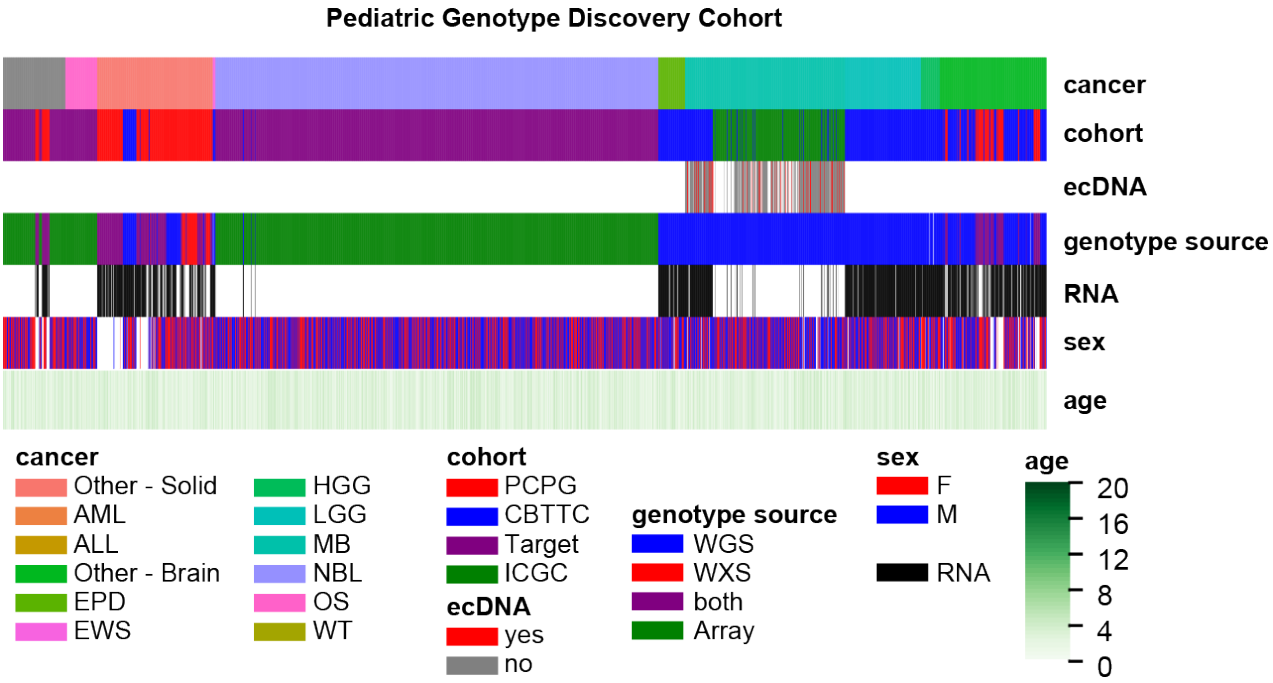

Figure 2.1 Pediatric Patient Characteristics Grid plot of pediatric population used for analyses. Cancer type, cohort (PCPG, CBTTC, Target, ICGC), ecDNA status, source of genotype information (WGS, WXS, Array), RNA availability, sex and age are indicated.

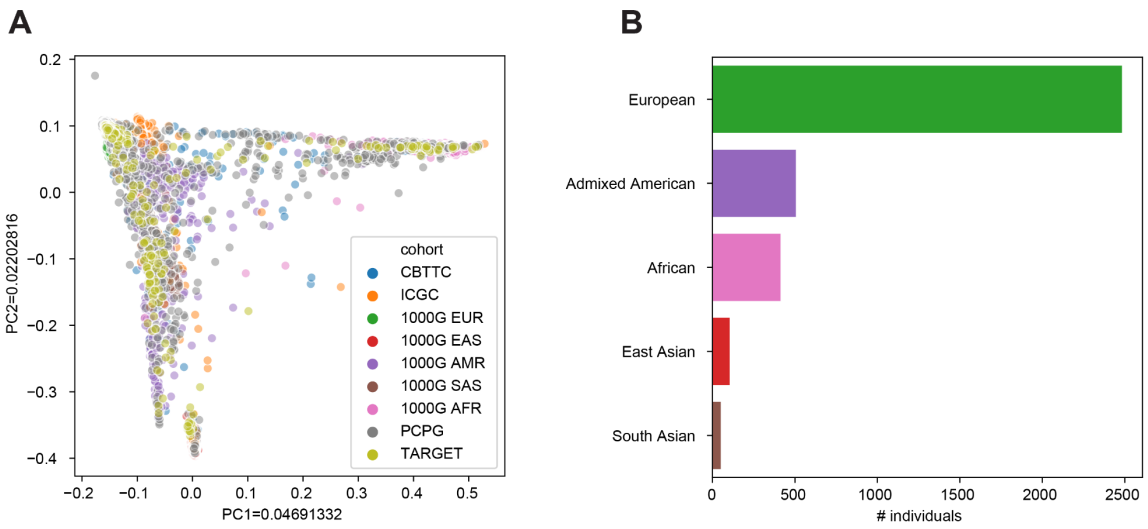

Figure 2.2 PCA of Pediatric Genotypes (A) PCA analysis of 547,600 intersecting germline variants; 1000 Genomes project and pediatric cohorts are highlighted in different colors. **(B)** Number of individuals in each ancestral group based on Sklearn clustering analysis.

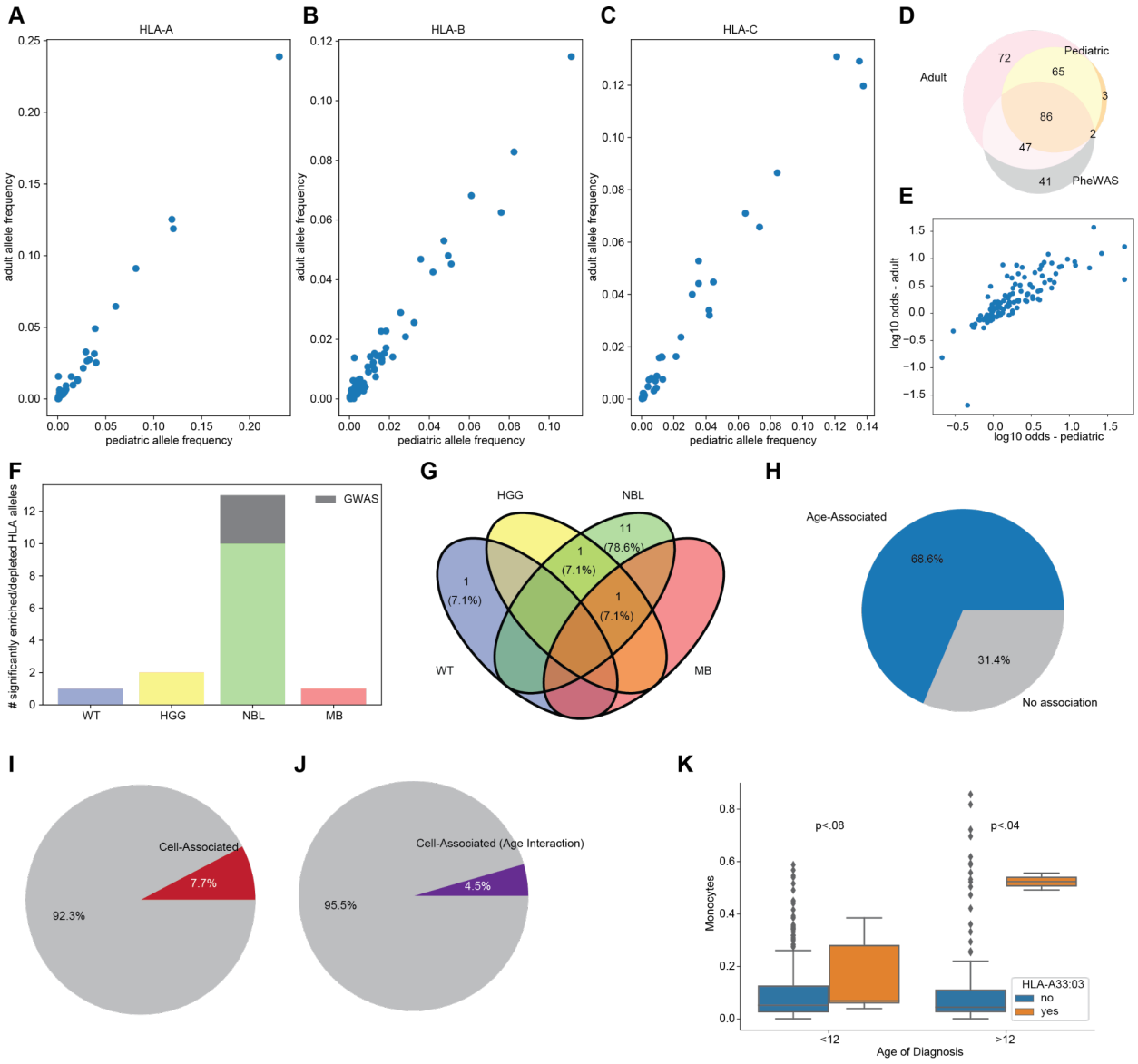

Figure 2.3 HLA Profiling of Pediatric Tumors Identified Alleles Specifically Associated with Pediatric Cancer and Age of Diagnosis Scatterplot of HLA-A **(A)**, HLA-B **(B)**, and HLA-C **(C)** allele frequencies in pediatric cohort and adult TCGA population (n=10428). **(D)** Overlap in signficant (FDR < .05) HLA alleles associated with pediatric and adult cancer compared to known cancer-risk HLA alleles from Vanderbilt PheWAS catalog (gray). **(E)** Comparison of Fisher's exact odds ratios between adult and pediatric cancer compared to noncancer control in UK Biobank. **(F)** Significant HLA alleles associated with each pediatric cancer type compared to UK Biobank noncancer controls. **(G)** Overlap between pediatric cancer-type specific HLA alleles **(H)** Proportion of pediatric cancer HLA alleles associated with age of diagnosis. **(I)** Proportion of pediatric cancer HLA alleles associated with immune cell infiltration. **(J)** Proportion of pediatric cancer HLA alleles associated with immune cell infiltration that interact with age. **(K)** HLA-A33:03 is associated with age of diagnosis of pediatric cancer and effects are age-dependent, with stronger effects observed in >12 years age group compared to <12 years age group.

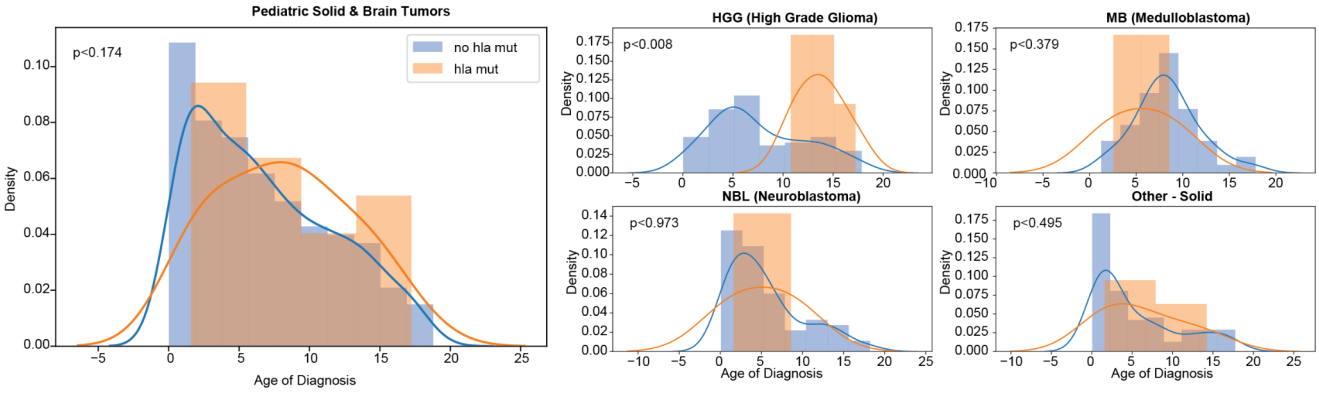

Figure 2.4 Age Distribution in PCPG Cohort Patients with Somatic HLA Mutations Density plot of age of diagnosis for pediatric cancer patients with or without a somatic HLA mutation. Cancer types for which > 1 individual has an HLA mutation are shown.

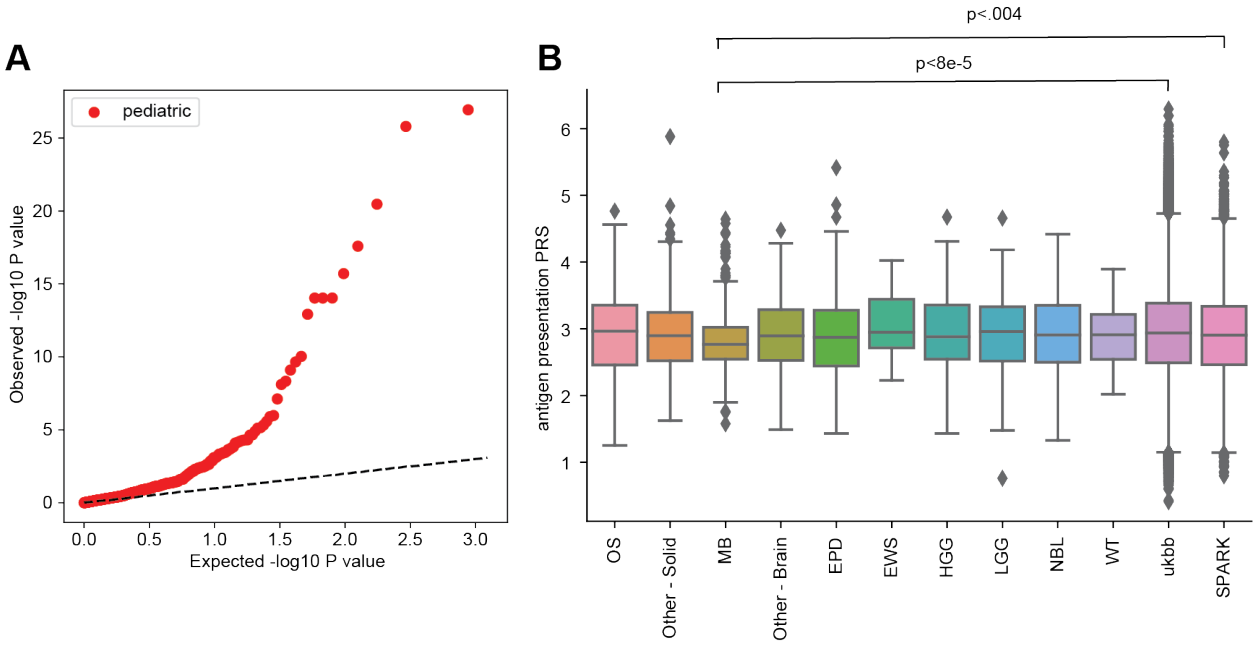

Figure 2.5 Adult Tumor Immune Microenvironment Variants Also Modify Pediatric Tumor Immune Microenvironments (A) QQ plot of adult TIME-SNPs in pediatric cancer cohort. **(B)** Distribution of melanoma risk antigen presentation PRS in pediatric cancers with controls of the non-cancer UK Biobank population and the SPARK autism cohort.

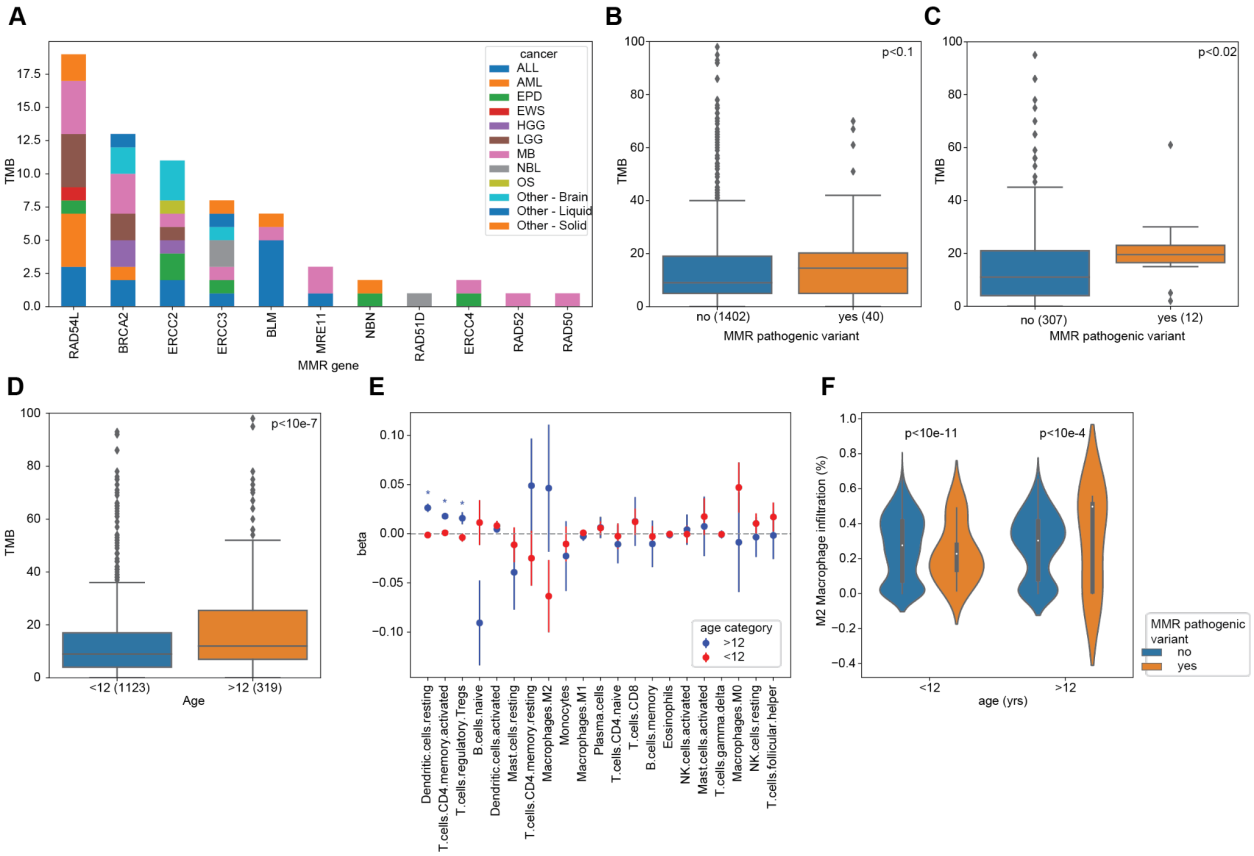

Figure 2.6 DNA Damage Repair (DDR) Pathogenic Variants and Effect on Pediatric Tumor Immune Microenvironment (A) Burden of DDR Pathogenic Variants in pediatric patients. **(B)** Boxplot of mutational load in pediatric patients with and without DDR pathogenic variant. **(C)** Boxplot of mutational load in medulloblastoma patients with and without DDR pathogenic variant. **(D)** Boxplot of mutational load in pediatric patients age < 12 and > 12 years. **(E)** Linear regression beta coefficient of DDR pathogenic variant with CIBERSORTx infiltration estimates in pediatric patients > 12 and < 12 years. **(F)** Violinplot of M2 Macrophage infiltration in pediatric patients with MMR pathogenic variant categorized by age.

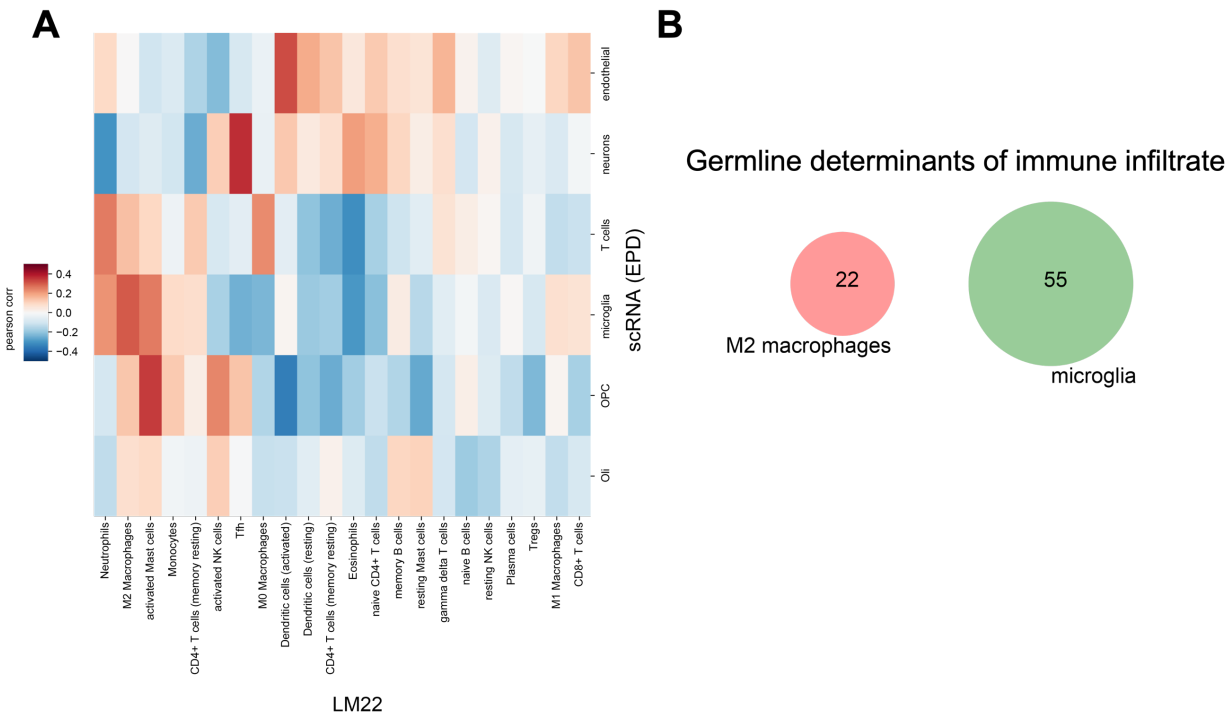

Figure 2.7 Comparison of Cell Infiltrates from scRNA-seq Ependymoma Signature Matrix and Default LM22 Matrix (A) Pearson correlation of immune cell infiltrate estimates using scRNA-seq ependymoma signature matrix and LM22 default matrix. **(B)** Venn diagram of overlap between germline variants associated with M2 macrophages from default LM22 matrix signature matrix and microglia from scRNA-seq ependymoma matrix.

2.1.7 Tables

Table 2.1 Pediatric Tumor Characteristics Distribution of pediatric tumors amongst 4 cohorts analyzed. ALL - acute lymphoid leukemia, AML - acute myeloid leukemia, NBL - neuroblastoma, WT - Wilm's Tumor, OS - osteosarcoma, HGG - high grade glioma, LGG - low grade glioma, EWS - Ewing's Sarcoma, MB - medulloblastoma, EPD - ependymoma, Other - solid includes sarcoma, rhabdoid tumors, GISTs, teratoma, Other - brain includes ganglioglioma, craniopharyngioma, Dysembryoplastic neuroepithelial tumor (DNET), Meningioma, Choroid plexus papilloma, Supratentorial or Spinal Cord PNET, Brainstem glioma- Diffuse intrinsic pontine glioma, Glial-neuronal tumor NOS, Schwannoma, Choroid plexus carcinoma, Subependymal Giant Cell Astrocytoma (SEGA), Malignant peripheral nerve sheath tumor (MPNST), Oligodendroglioma, Neurofibroma/plexiform, germinoma, chordoma. Patients with dysplasia/gliosis, other, not available, tuberous sclerosis, langerhans cell histiocytosis, NF-1, NF-2, Holt Oram syndrome, adenoma, cavernoma, schwannomatosis, gliomatosis cerebri, embryonal tumor with multilayerd rosettes were excluded as they represented benign growths or did not specify a clear tumor type category.

cohort	cancer	# patients
Target	ALL	1248
Target	NBL	1104
Target	AML	355
Target	WT	126
Target	OS	80
Target	Other - solid	53
PCPG	ALL	464
PCPG	BALL	322
PCPG	AML	284
PCPG	Other - Solid	251
PCPG	HGG	153
PCPG	NBL	150
PCPG	LGG	92
PCPG	Other - Brain	91
PCPG	MB	90
PCPG	EPD	77
PCPG	OS	62
PCPG	EWS	32
PCPG	WT	29
PCPG	Other - Liquid	25
ICGC	MB	304
CBTTC	LGG	190
CBTTC	Other - Brain	176
CBTTC	MB	96
CBTTC	EPD	67
CBTTC	HGG	48
CBTTC	Other - Solid	38
CBTTC	EWS	7
CBTTC	NBL	4

2.1.8 Supplemental Figures

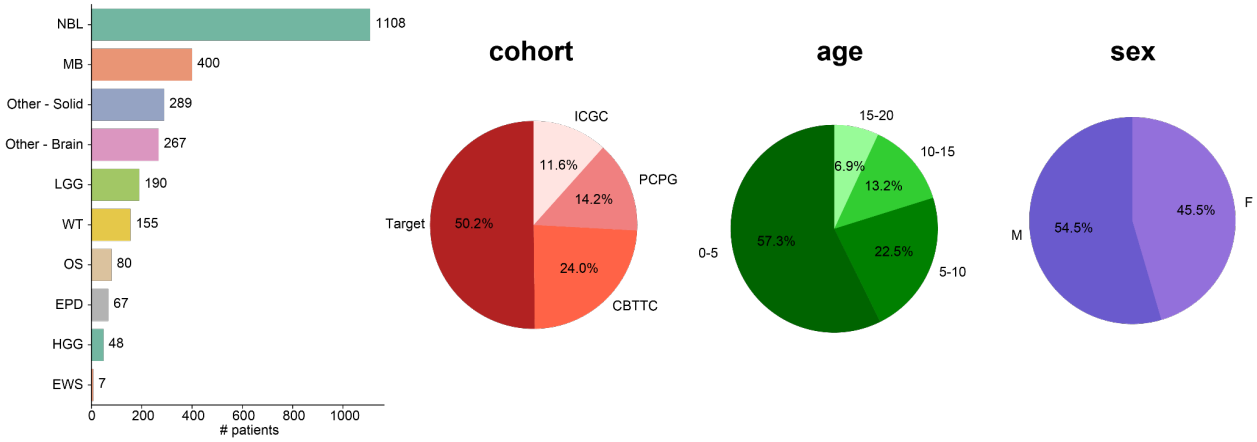

Figure S2.1 Pediatric Patient Characteristics Barplot of pediatric cancer types used for analysis. Piecharts of cohort, age and sex of pediatric cancers are also given.

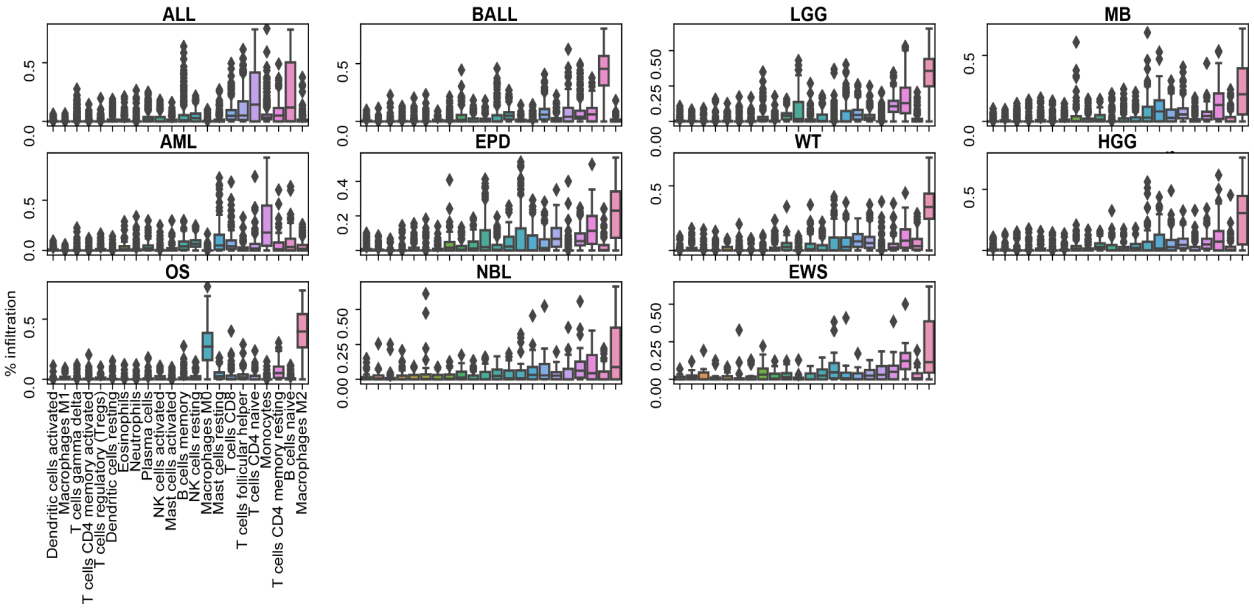

Figure S2.2 CIBERSORTx Infiltration by Pediatric Cancer Type Infiltration estimates of 22 immune cell subsets based on CIBERSORTx LM22 signature matrix in pediatric tumors divided by cancer type.

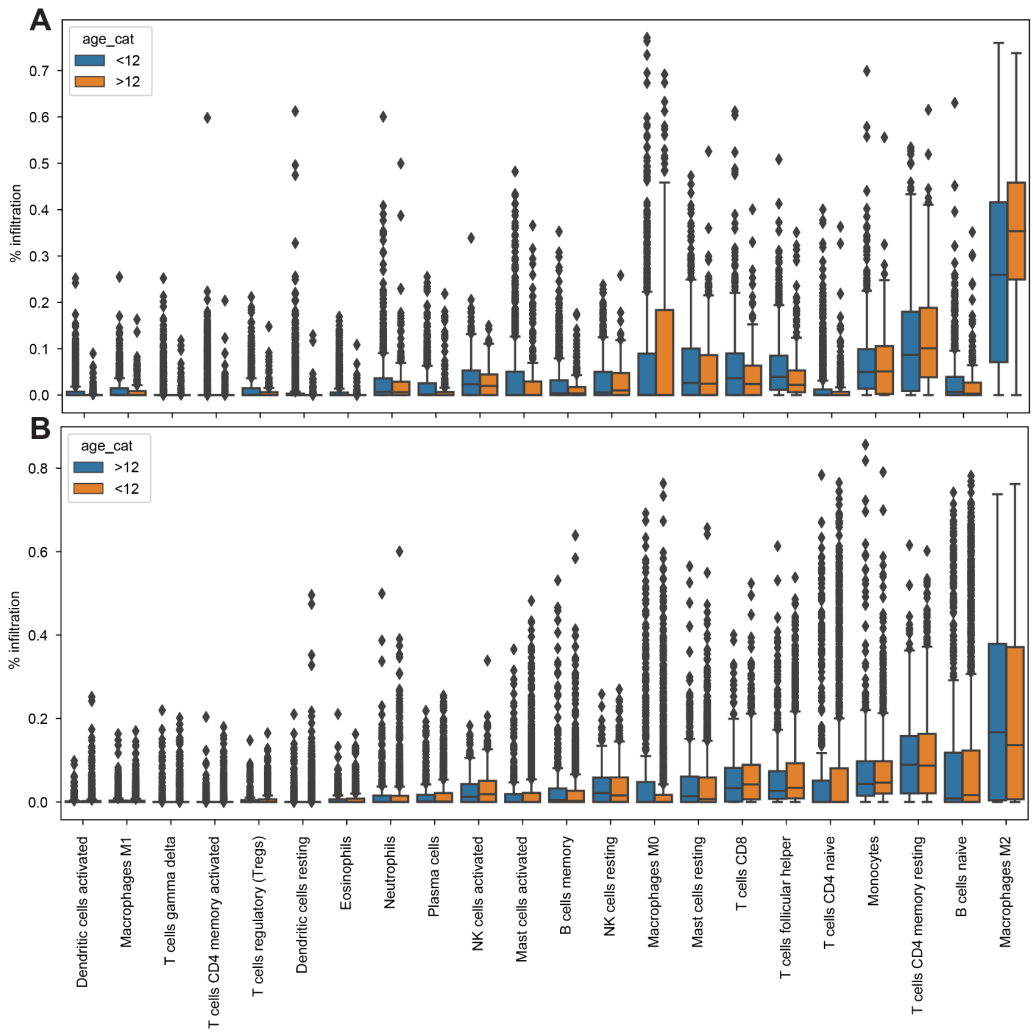

Figure S2.3 CIBERSORTx Infiltration in Pediatric Cancers Divided by Older and Younger Pediatric Patients Infiltration estimates of 22 immune cell subsets based on CIBERSORTx LM22 signature matrix in pediatric tumors divided by age subset in (A) solid tumors and (B) all tumors.

2.1.9 Author Contributions

Original concept: Meghana Pagadala, Hannah Carter, Jill Mesirov

Data analysis: Meghana Pagadala, Owen Chapman, Lukas Chavez, Andrea Castro

UK Biobank genetic analysis: Meghana Pagadala

Manuscript writing: Meghana Pagadala, Hannah Carter, Jill Mesirov, Rob Wechsler-Reya

2.1.10 Acknowledgements

The results published here are in whole or part based upon data generated by the Therapeutically Applicable Research to Generate Effective Treatments (https://ocg.cancer.gov/programs/target) initiative, phs000218. The data used for this analysis are available at https://portal.gdc.cancer.gov/projects.

Chapter 2, in part, is a reformatted reprint of material currently being prepared for submission for publication of the material. Pagadala; Meghana; Castro, Andrea; Sears, Timothy; Chavez, Lukas; Chapman, Owen; Wechsler-Reya, Robert; Mesirov, Jill*; Carter, Hannah*. The dissertation author was the primary author of this material.

2.1.11 References

1. R. M. Cunningham, M. A. Walton, P. M. Carter, The Major Causes of Death in Children and Adolescents in the United States. *New England Journal of Medicine*. **379** (2018), pp. 2468–2475.

2. R. L. Siegel, K. D. Miller, A. Jemal, Cancer statistics, 2016. *CA Cancer J. Clin.* **66**, 7–30 (2016).

3. L. A. Gloeckler Ries, *Cancer Incidence and Survival Among Children and Adolescents: United States SEER Program, 1975-1995* (1999).

4. A. B. Ryerson, C. R. Eheman, S. F. Altekruse, J. W. Ward, A. Jemal, R. L. Sherman, S. J. Henley, D. Holtzman, A. Lake, A.-M. Noone, R. N. Anderson, J. Ma, K. N. Ly, K. A. Cronin, L. Penberthy, B. A. Kohler, Annual Report to the Nation on the Status of Cancer, 1975-2012, featuring the increasing incidence of liver cancer. *Cancer*. **122**, 1312–1337 (2016).

5. M. Grunert, R. Kassubek, B. Danz, B. Klemenz, S. Hasslacher, S. Stroh, L. Schneele, J. Langhans, S. Ströbele, S. E. Barry, S. Zhou, K.-M. Debatin, M.-A. Westhoff, Radiation and Brain Tumors: An Overview. *Crit. Rev. Oncog.* **23**, 119–138 (2018).

6. R. L. Siegel, K. D. Miller, H. E. Fuchs, A. Jemal, Cancer Statistics, 2021. *CA Cancer J. Clin.* **71**, 7–33 (2021).

7. B. B. Campbell, N. Light, D. Fabrizio, M. Zatzman, F. Fuligni, R. de Borja, S. Davidson, M. Edwards, J. A. Elvin, K. P. Hodel, W. J. Zahurancik, Z. Suo, T. Lipman, K. Wimmer, C. P. Kratz, D. C. Bowers, T. W. Laetsch, G. P. Dunn, T. M. Johanns, M. R. Grimmer, I. V. Smirnov, V. Larouche, D. Samuel, A. Bronsema, M. Osborn, D. Stearns, P. Raman, K. A. Cole, P. B. Storm, M. Yalon, E. Opocher, G. Mason, G. A. Thomas, M. Sabel, B. George, D. S. Ziegler, S. Lindhorst, V. M. Issai, S. Constantini, H. Toledano, R. Elhasid, R. Farah, R. Dvir, P. Dirks, A. Huang, M. A. Galati, J. Chung, V. Ramaswamy, M. S. Irwin, M. Aronson, C. Durno, M. D. Taylor, G. Rechavi, J. M. Maris, E. Bouffet, C. Hawkins, J. F. Costello, M. S. Meyn, Z. F. Pursell, D. Malkin, U. Tabori, A. Shlien, Comprehensive Analysis of Hypermutation in Human Cancer. *Cell*. **171**, 1042–1056.e10 (2017).

8. E. A. Sweet-Cordero, J. A. Biegel, The genomic landscape of pediatric cancers: Implications for diagnosis and treatment. *Science*. **363**, 1170–1175 (2019).

9. X. Ma, Y. Liu, Y. Liu, L. B. Alexandrov, M. N. Edmonson, C. Gawad, X. Zhou, Y. Li, M. C. Rusch, J. Easton, R. Huether, V. Gonzalez-Pena, M. R. Wilkinson, L. C. Hermida, S. Davis, E. Sioson, S. Pounds, X. Cao, R. E. Ries, Z. Wang, X. Chen, L. Dong, S. J. Diskin, M. A. Smith, J. M. Guidry Auvil, P. S. Meltzer, C. C. Lau, E. J. Perlman, J. M. Maris, S. Meshinchi, S. P. Hunger, D. S. Gerhard, J. Zhang, Pan-cancer genome and transcriptome analyses of 1,699 paediatric leukaemias and solid tumours. *Nature*. **555**, 371–376 (2018).

10. B. S. Paugh, C. Qu, C. Jones, Z. Liu, M. Adamowicz-Brice, J. Zhang, D. A. Bax, B. Coyle, J. Barrow, D. Hargrave, J. Lowe, A. Gajjar, W. Zhao, A. Broniscer, D. W. Ellison, R. G.

Grundy, S. J. Baker, Integrated molecular genetic profiling of pediatric high-grade gliomas reveals key differences with the adult disease. *J. Clin. Oncol.* **28**, 3061–3068 (2010).

11. F. Petralia, N. Tignor, B. Reva, M. Koptyra, S. Chowdhury, D. Rykunov, A. Krek, W. Ma, Y. Zhu, J. Ji, A. Calinawan, J. R. Whiteaker, A. Colaprico, V. Stathias, T. Omelchenko, X. Song, P. Raman, Y. Guo, M. A. Brown, R. G. Ivey, J. Szpyt, S. Guha Thakurta, M. A. Gritsenko, K. K. Weitz, G. Lopez, S. Kalayci, Z. H. Gümüş, S. Yoo, F. da Veiga Leprevost, H.-Y. Chang, K. Krug, L. Katsnelson, Y. Wang, J. J. Kennedy, U. J. Voytovich, L. Zhao, K. S. Gaonkar, B. M. Ennis, B. Zhang, V. Baubet, L. Tauhid, J. V. Lilly, J. L. Mason, B. Farrow, N. Young, S. Leary, J. Moon, V. A. Petyuk, J. Nazarian, N. D. Adappa, J. N. Palmer, R. M. Lober, S. Rivero-Hinojosa, L.-B. Wang, J. M. Wang, M. Broberg, R. K. Chu, R. J. Moore, M. E. Monroe, R. Zhao, R. D. Smith, J. Zhu, A. I. Robles, M. Mesri, E. Boja, T. Hiltke, H. Rodriguez, B. Zhang, E. E. Schadt, D. R. Mani, L. Ding, A. Iavarone, M. Wiznerowicz, S. Schürer, X. S. Chen, A. P. Heath, J. L. Rokita, A. I. Nesvizhskii, D. Fenyö, K. D. Rodland, T. Liu, S. P. Gygi, A. G. Paulovich, A. C. Resnick, P. B. Storm, B. R. Rood, P. Wang, Children's Brain Tumor Network, Clinical Proteomic Tumor Analysis Consortium, Integrated Proteogenomic Characterization across Major Histological Types of Pediatric Brain Cancer. *Cell*. **183**, 1962–1985.e31 (2020).

12. B. Geoerger, H. J. Kang, M. Yalon-Oren, L. V. Marshall, C. Vezina, A. Pappo, T. W. Laetsch, A. S. Petrilli, M. Ebinger, J. Toporski, J. Glade-Bender, W. Nicholls, E. Fox, S. G. DuBois, M. E. Macy, S. L. Cohn, K. Pathiraja, S. J. Diede, S. Ebbinghaus, N. Pinto, Pembrolizumab in paediatric patients with advanced melanoma or a PD-L1-positive, advanced, relapsed, or refractory solid tumour or lymphoma (KEYNOTE-051): interim analysis of an open-label, single-arm, phase 1-2 trial. *Lancet Oncol.* **21**, 121–133 (2020).

13. B. Geoerger, C. Michel Zwaan, L. V. Marshall, J. Michon, F. Bourdeaut, M. Casanova, N. Corradini, G. Rossato, M. Farid-Kapadia, C. S. Shemesh, K. E. Hutchinson, F. Donaldson, M. Liao, H. Caron, T. Trippett, Atezolizumab for children and young adults with previously treated solid tumours, non-Hodgkin lymphoma, and Hodgkin lymphoma (iMATRIX): a multicentre phase 1–2 study. *The Lancet Oncology*. **21** (2020), pp. 134–144.

14. M. S. Merchant, M. Wright, K. Baird, L. H. Wexler, C. Rodriguez-Galindo, D. Bernstein, C. Delbrook, M. Lodish, R. Bishop, J. D. Wolchok, H. Streicher, C. L. Mackall, Phase I Clinical Trial of Ipilimumab in Pediatric Patients with Advanced Solid Tumors. *Clin. Cancer Res.* **22**, 1364–1370 (2016).

15. K. L. Davis, E. Fox, M. S. Merchant, J. M. Reid, R. A. Kudgus, X. Liu, C. G. Minard, S. Voss, S. L. Berg, B. J. Weigel, C. L. Mackall, Nivolumab in children and young adults with relapsed or refractory solid tumours or lymphoma (ADVL1412): a multicentre, open-label, single-arm, phase 1-2 trial. *Lancet Oncol.* **21**, 541–550 (2020).

16. R. Marty, S. Kaabinejadian, D. Rossell, M. J. Slifker, J. van de Haar, H. B. Engin, N. de Prisco, T. Ideker, W. H. Hildebrand, J. Font-Burgada, H. Carter, MHC-I Genotype Restricts the Oncogenic Mutational Landscape. *Cell*. **171**, 1272–1283.e15 (2017).

17. R. Marty Pyke, W. K. Thompson, R. M. Salem, J. Font-Burgada, M. Zanetti, H. Carter,

Evolutionary Pressure against MHC Class II Binding Cancer Mutations. *Cell*. **175**, 1991 (2018).

18. A. M. Goodman, A. Castro, R. M. Pyke, R. Okamura, S. Kato, P. Riviere, G. Frampton, E. Sokol, X. Zhang, E. D. Ball, H. Carter, R. Kurzrock, MHC-I genotype and tumor mutational burden predict response to immunotherapy. *Genome Med.* **12**, 45 (2020).

19. T.-T. Sheu, B.-L. Chiang, J.-H. Yen, W.-C. Lin, Premature CD4+ T cell aging and its contribution to lymphopenia-induced proliferation of memory cells in autoimmune-prone non-obese diabetic mice. *PLoS One*. **9**, e89379 (2014).

20. W. T. Shearer, H. M. Rosenblatt, R. S. Gelman, R. Oyomopito, S. Plaeger, E. Richard Stiehm, D. W. Wara, S. D. Douglas, K. Luzuriaga, E. J. McFarland, R. Yogev, M. H. Rathore, W. Levy, B. L. Graham, S. A. Spector, Lymphocyte subsets in healthy children from birth through 18 years of age. *Journal of Allergy and Clinical Immunology*. **112** (2003), pp. 973–980.

21. D. Ran, M. Cai, X. Zhang, Genetics of psoriasis: a basis for precision medicine. *Precision Clinical Medicine*. **2** (2019), pp. 120–130.

22. SPARK Consortium. Electronic address: pfeliciano@simonsfoundation.org, SPARK Consortium, SPARK: A US Cohort of 50,000 Families to Accelerate Autism Research. *Neuron*. **97**, 488–493 (2018).

23. A. Shlien, B. B. Campbell, R. de Borja, L. B. Alexandrov, D. Merico, D. Wedge, P. Van Loo, P. S. Tarpey, P. Coupland, S. Behjati, A. Pollett, T. Lipman, A. Heidari, S. Deshmukh, N. 'ama Avitzur, B. Meier, M. Gerstung, Y. Hong, D. M. Merino, M. Ramakrishna, M. Remke, R. Arnold, G. B. Panigrahi, N. P. Thakkar, K. P. Hodel, E. E. Henninger, A. Y. Göksenin, D. Bakry, G. S. Charames, H. Druker, J. Lerner-Ellis, M. Mistry, R. Dvir, R. Grant, R. Elhasid, R. Farah, G. P. Taylor, P. C. Nathan, S. Alexander, S. Ben-Shachar, S. C. Ling, S. Gallinger, S. Constantini, P. Dirks, A. Huang, S. W. Scherer, R. G. Grundy, C. Durno, M. Aronson, A. Gartner, M. S. Meyn, M. D. Taylor, Z. F. Pursell, C. E. Pearson, D. Malkin, P. A. Futreal, M. R. Stratton, E. Bouffet, C. Hawkins, P. J. Campbell, U. Tabori, Biallelic Mismatch Repair Deficiency Consortium, Combined hereditary and somatic mutations of replication error repair genes result in rapid onset of ultra-hypermutated cancers. *Nat. Genet.* **47**, 257–262 (2015).

24. A. Das, S. Sudhaman, D. Morgenstern, A. Coblentz, J. Chung, S. C. Stone, N. Alsafwani, Z. A. Liu, O. A. A. Karsaneh, S. Soleimani, H. Ladany, D. Chen, M. Zatzman, V. Cabric, L. Nobre, V. Bianchi, M. Edwards, L. C. Sambira Nahum, A. B. Ercan, A. Nabbi, S. Constantini, R. Dvir, M. Yalon-Oren, G. A. Campino, S. Caspi, V. Larouche, A. Reddy, M. Osborn, G. Mason, S. Lindhorst, A. Bronsema, V. Magimairajan, E. Opocher, R. L. De Mola, M. Sabel, C. Frojd, D. Sumerauer, D. Samuel, K. Cole, S. Chiaravalli, M. Massimino, P. Tomboc, D. S. Ziegler, B. George, A. Van Damme, N. Hijiya, D. Gass, R. B. McGee, O. Mordechai, D. C. Bowers, T. W. Laetsch, A. Lossos, D. T. Blumenthal, T. Sarosiek, L. Y. Yen, J. Knipstein, A. Bendel, L. M. Hoffman, S. Luna-Fineman, S. Zimmermann, I. Scheers, K. E. Nichols, M. Zapotocky, J. R. Hansford, J. M. Maris, P. Dirks, M. D. Taylor, A. V. Kulkarni, M. Shroff,

D. S. Tsang, A. Villani, W. Xu, M. Aronson, C. Durno, A. Shlien, D. Malkin, G. Getz, Y. E. Maruvka, P. S. Ohashi, C. Hawkins, T. J. Pugh, E. Bouffet, U. Tabori, Genomic predictors of response to PD-1 inhibition in children with germline DNA replication repair deficiency. *Nat. Med.* **28**, 125–135 (2022).

25. D. Hsiehchen, A. Hsieh, R. M. Samstein, T. Lu, M. S. Beg, D. E. Gerber, T. Wang, L. G. T. Morris, H. Zhu, DNA Repair Gene Mutations as Predictors of Immune Checkpoint Inhibitor Response beyond Tumor Mutation Burden. *Cell Rep Med*. **1** (2020), doi:10.1016/j.xcrm.2020.100034.

26. A. M. Newman, C. L. Liu, M. R. Green, A. J. Gentles, W. Feng, Y. Xu, C. D. Hoang, M. Diehn, A. A. Alizadeh, Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods*. **12**, 453–457 (2015).

27. J. Gojo, B. Englinger, L. Jiang, J. M. Hübner, M. L. Shaw, O. A. Hack, S. Madlener, D. Kirchhofer, I. Liu, J. Pyrdol, V. Hovestadt, E. Mazzola, N. D. Mathewson, M. Trissal, D. Lötsch, C. Dorfer, C. Haberler, A. Halfmann, L. Mayr, A. Peyrl, R. Geyeregger, B. Schwalm, M. Mauermann, K. W. Pajtler, T. Milde, M. E. Shore, J. E. Geduldig, K. Pelton, T. Czech, O. Ashenberg, K. W. Wucherpfennig, O. Rozenblatt-Rosen, S. Alexandrescu, K. L. Ligon, S. M. Pfister, A. Regev, I. Slavc, W. Berger, M. L. Suvà, M. Kool, M. G. Filbin, Single-Cell RNA-Seq Reveals Cellular Hierarchies and Impaired Developmental Trajectories in Pediatric Ependymoma. *Cancer Cell*. **38**, 44–59.e9 (2020).

28. J. Vakkila, R. Jaffe, M. Michelow, M. T. Lotze, Pediatric Cancers Are Infiltrated Predominantly by Macrophages and Contain a Paucity of Dendritic Cells: a Major Nosologic Difference with Adult Tumors. *Clinical Cancer Research*. **12** (2006), pp. 2049–2054.

29. T. Aoki, M. Hino, K. Koh, M. Kyushiki, H. Kishimoto, Y. Arakawa, R. Hanada, H. Kawashima, J. Kurihara, N. Shimojo, S. Motohashi, Low Frequency of Programmed Death Ligand 1 Expression in Pediatric Cancers. *Pediatr. Blood Cancer*. **63**, 1461–1464 (2016).

30. N. Pinto, J. R. Park, E. Murphy, J. Yearley, T. McClanahan, L. Annamalai, D. S. Hawkins, E. R. Rudzinski, Patterns of PD-1, PD-L1, and PD-L2 expression in pediatric solid tumors. *Pediatr. Blood Cancer*. **64** (2017), doi:10.1002/pbc.26613.

31. C. McLeod, A. M. Gout, X. Zhou, A. Thrasher, D. Rahbarinia, S. W. Brady, M. Macias, K. Birch, D. Finkelstein, J. Sunny, R. Mudunuri, B. A. Orr, M. Treadway, B. Davidson, T. K. Ard, A. Chiao, A. Swistak, S. Wiggins, S. Foy, J. Wang, E. Sioson, S. Wang, J. R. Michael, Y. Liu, X. Ma, A. Patel, M. N. Edmonson, M. R. Wilkinson, A. M. Frantz, T.-C. Chang, L. Tian, S. Lei, S. M. A. Islam, C. Meyer, N. Thangaraj, P. Tater, V. Kandali, S. Ma, T. Nguyen, O. Serang, I. McGuire, N. Robison, D. Gentry, X. Tang, L. E. Palmer, G. Wu, E. Suh, L. Tanner, J. McMurry, M. Lear, A. S. Pappo, Z. Wang, C. L. Wilson, Y. Cheng, S. Meshinchi, L. B. Alexandrov, M. J. Weiss, G. T. Armstrong, L. L. Robison, Y. Yasui, K. E. Nichols, D. W. Ellison, C. Bangur, C. G. Mullighan, S. J. Baker, M. A. Dyer, G. Miller, S. Newman, M. Rusch, R. Daly, K. Perry, J. R. Downing, J. Zhang, St. Jude Cloud: A Pediatric Cancer Genomic Data-Sharing Ecosystem. *Cancer Discov.* **11**, 1082–1099 (2021).

32. J. R. Downing, R. K. Wilson, J. Zhang, E. R. Mardis, C.-H. Pui, L. Ding, T. J. Ley, W. E.

Evans, The Pediatric Cancer Genome Project. *Nat. Genet.* **44**, 619–622 (2012).

33. S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, P. C. Sham, PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*. **81** (2007), pp. 559–575.

34. S. A. Shukla, M. S. Rooney, M. Rajasagi, G. Tiao, P. M. Dixon, M. S. Lawrence, J. Stevens, W. J. Lane, J. L. Dellagatta, S. Steelman, C. Sougnez, K. Cibulskis, A. Kiezun, N. Hacohen, V. Brusic, C. J. Wu, G. Getz, Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. *Nature Biotechnology*. **33** (2015), pp. 1152–1158.

35. A. M. Newman, C. B. Steen, C. L. Liu, A. J. Gentles, A. A. Chaudhuri, F. Scherer, M. S. Khodadoust, M. S. Esfahani, B. A. Luca, D. Steiner, M. Diehn, A. A. Alizadeh, Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat. Biotechnol.* **37**, 773–782 (2019).

CHAPTER 3: Germline Analysis of Androgens and Prostate Cancer in Diverse Ancestral Cohort

3.1.1 Abstract

Utilizing data from the Million Veteran Program (MVP), we investigated the genetic determinants underlying total testosterone levels via a multi-ancestral analysis of 124,593 individuals of European (n=88,385), African (n=25,235) and Hispanic (n=10,973) ancestry. We identified 46 trans-ancestry variants and 17 ancestry-specific variants, of which 14 trans-ancestry variants and 15 ancestry-specific variants are novel associations with testosterone. Results implicate genes regulating testosterone shared across ancestral groups, which include *SHBG, JMJD1C, FXR2, SENP3, TNFSF12-TNFSF13* while also implicating genes such as *MSN, DMD, VSIG4, CHEK2, TKTL1* that may underlie ancestry-group differences in testosterone regulation. We also linked testosterone variants on the X chromosome with differential risk of chronic kidney disease and hereditary hemolytic anemias in African and Hispanic ancestry groups, respectively. Lastly, we constructed a polygenic score from our 46 trans-ancestry variants and associated it with testicular dysfunction, hyperlipidemia, gout and prostate cancer with stronger prostate cancer associations in Hispanic and African ancestry groups compared to the European ancestry group. These findings provide insight into ancestry-specific androgen regulation and identify novel variants for disease risk stratification in patients.

3.1.2 Introduction

Testosterone is an anabolic steroid synthesized by the testis which acts as the primary sex hormone in men regulating sexual development during puberty and spermatogenesis and sexual function in adulthood(*1–3*). Classical, well-established roles of testosterone include stimulation of

erythropoiesis, maintenance of muscular strength, and maintenance of bone density mass;(*4*, *5*) however, research has discovered that testicular androgens, testosterone and dihydrotestosterone have more extensive physiological actions with important roles in regulating cardiovascular, metabolic, hepatic, immune, and brain function(*6–8*). These ubiquitous actions of testosterone are mediated by the androgen receptor (AR), encoded by the AR gene on the X chromosome, which is widely expressed in the testes, prostate, bone, skeletal muscle, heart, vascular smooth muscle, kidney, pulmonary epithelial cells, bone marrow hematopoietic and immune cells, adipose tissue, and the central nervous system(*2*, *9*).

Beginning in midlife men experience a progressive reduction in testicular androgen steroidogenesis, which results in a roughly 1-3% decline in circulating levels of total testosterone per year(*1*, *2*, *10*, *11*). Many men develop an age-related testosterone deficiency or hypogonadism, which contributes to sexual dysfunction, frailty, sarcopenia, coronary artery disease, metabolic syndrome and type 2 diabetes(*1*, *2*, *10–12*). Testosterone levels in the hypogonadal range have been reported to increase the risk of prostate cancer and predict recurrence after prostatectomy and poor outcome in metastatic prostate cancer(*13*). Testosterone deficiency has additionally been linked to depression, neurodegeneration, cognitive impairment, and Alzheimer's disease risk during male aging(*14–18*). To date, the genetic factors that contribute to individual differences in circulating testosterone levels in men during aging remain poorly understood. Furthermore, analysis of ancestry-specific genetic determinants governing testosterone regulation has not been studied.

Total testosterone levels are heritable (~20%) and a number of genome-wide association studies (GWAS) of testosterone levels have been conducted, identifying strong associations with *SHBG, JMJD1C, FKBP4, REEP3,* and *FAM9B(19–25)*. However, most GWAS studies have been

conducted primarily in European or Asian ancestry groups whereas differences in testosterone levels have been observed in men of other ancestries(*26–28*). In this study, we conducted a GWAS of endogenous total testosterone levels in men enrolled in the Million Veteran Program (MVP), a large, multi-ethnic genetic biorepository from a hospital-based population. As a hospital-based population, MVP has higher disease prevalence of type 2 diabetes, obesity, hyperlipidemia, sleep apnea, gout, prostate cancer, and testicular dysfunction/hypofunction compared to other population studies and offers unique insights into testosterone regulation in a clinical setting. We identified 17 ancestry-specific testosterone gene variants and 46 trans-ancestry testosterone gene variants, of which 15 ancestry-specific variants and 14 trans-ancestry variants were novel. Colocalization and transcriptome-wide association study (TWAS) analysis revealed novel cell-type specific associations between *BRI3* and *BAIAP2L1*. Finally, we conducted phenome (PheWAS) and laboratory value association studies (LabWAS), which revealed shared associations of testosterone with type 2 diabetes, hyperlipidemia, gout and liver disease and ancestry-specific association with diabetes in the African ancestry group and hemolytic conditions in the Hispanic ancestry group. With this study, we gain insight into ancestry-specific regulation of testosterone, highlighting the importance of conducting future studies in diverse ancestry groups.

3.1.3 Results

3.1.3.1 Genome-Wide Analysis of Testosterone Levels

After phenotyping quality control, morning total testosterone levels were available for 145,576 male MVP participants (**Figure S3.1)**. MVP participants were categorized using Harmonized Ancestry and Race/Ethnicity (HARE), an algorithm which uses both self-reported

and genetic ancestry to assign individuals to ancestry groups(*29*) **(Table 3.1)**. Genome-wide association studies (GWAS) with testosterone levels were conducted separately in European (n=88,385), African (n=25,235) and Hispanic (n=10,973) ancestry groups. Incorporating all genetic information, we first assessed the heritability of testosterone levels from autosomal chromosomes across ancestry groups using LD score regression(*30*). Total testosterone levels showed the highest heritability in the African ancestry group (h2=0.095) with significant GWAS hits (p<5e-08) accounting for 3.3% of heritability. Heritability of total testosterone levels in the European and Hispanic ancestry groups were 0.071 and 0.084 with significant hits (p<5e-08) accounting for 14.6% and 11.1% of heritability respectively **(Table 3.2)**. Variants passing the suggestive threshold (p<1e-05) explained 24.6%, 9.1%, 30.4% of heritability in European, African, and Hispanic ancestry groups, respectively.

GWAS analyses identified 63 significant (p<5e-08) variants associated with testosterone across ancestry groups: 30 significant autosomal loci for the European ancestry group, 5 for the African ancestry group and 3 for the Hispanic ancestry group (**Figure 3.1)**. Several X chromosome variants remained significant after linkage-based clumping. To identify independent loci, we conducted a conditional analysis, where we used the most significant variants as covariates in the GWAS. Conditional analysis identified 9, 10 and 6 associations in European, African and Hispanic cohorts, respectively **(Figure 3.2)**. Variants near *FAM9B, AR, RGAG1* were shared amongst all ancestry groups. X chromosome variant near *DMD* was identified specifically in the African ancestry group **(Figure 3.2)**. *DMD* is responsible for dystrophin protein production and is mainly expressed in skeletal and cardiac muscle, although some expression in the brain has been observed(*31–34*). A Hispanic ancestry group-specific variant near the ribosomal pseudogene *RPS29P28* was also identified **(Figure 3.2)**.

Loci on chromosomes 10 and 17 were shared across different ancestry groups. Most unique peaks were observed for the European ancestry group which had the largest sample size; however, unique peaks on chromosomes 3, 7, 8 and X were observed in the African ancestry group. We compared beta coefficients of SNP associations to determine if effects on testosterone were different amongst ancestry groups. Of the 63 variants, 10 variants had different effects on testosterone levels in the African ancestry group compared to the European ancestry group. Four of these variants were located on the X chromosome, with other variants located on chromosome 7, 8, 17, 19, and 22. Of the 63 variants, 8 variants had different effects on testosterone in the Hispanic ancestry group compared to the European ancestry group. Three of these variants were located on chromosome X, while the others were located on chromosomes 3, 7, 8, and 16. The chromosome 17 variant, rs980068134, was identified as a trans-ancestry variant and was associated with lower testosterone levels in African and Hispanic ancestry groups, but higher testosterone levels in the European ancestry group.

To validate the GWAS significant variants, we performed replication analyses for the European ancestry group variants in the UK Biobank. Of the 39 European ancestry group variants, 31 passed a significance threshold of 5e-08 in the UK Biobank (n=166,502) **(Figure S3.2)**. Of the remaining 8 which did not pass GWAS significance threshold, 5 passed a significance threshold of 0.05. Comparable cohorts for the African and Hispanic ancestry group powered for replication were unavailable. However, we obtained total testosterone levels for individuals of African ancestry (n=1474) in the Multi-Ethnic Study of Atherosclerosis (MESA) study and found beta values of variants were concordant. **(Figure S3.2)**.

3.1.3.2 Genetic Correlation Across Ancestry Groups

To evaluate the extent to which variants modifying total testosterone levels are shared between ancestry groups, we estimated genetic correlation between groups by LD score (LDSC) regression(*35*). While there was significant correlation between all groups, higher genetic correlation was observed between European and Hispanic ancestry groups (rg=0.60 [0.42-0.78]) compared to European and African ancestry groups (rg=0.19 [0.12-0.26]). Genetic correlation between African and Hispanic ancestry groups was 0.37 [0.16-0.58] **(Figure 3.3)**.

We next conducted genetic correlation analysis between our total testosterone GWAS results and diseases/traits profiled in previous published GWASs using LDhub(*36*), assessing differences in ancestry groups. Significant positive correlation with HDL and lung function, including forced vital capacity (FVC) and forced expiratory volume (FEV), were observed in European ancestry group (FDR < 0.1). Total testosterone levels were significantly negatively correlated with obesity, BMI, Type 2 Diabetes, triglycerides and VLDL (FDR < 0.1) in the European ancestry group. **(Figure 3.3).** No genetic correlations were significant in African and Hispanic ancestry groups; however, genetic correlations were highly concordant amongst ancestry groups **(Figure S3.3)**. Positive correlations between testosterone and glucose were observed in the Hispanic ancestry group, but not European and African ancestry groups (**Figure S3.3)**. These findings suggest potential differences in testosterone regulation of glucose between ancestry groups.

3.1.3.3 Testosterone Variants Affect Gene Expression in Specific Cell Types

In order to identify candidate genes influencing testosterone levels we conducted a colocalization and a transcriptome-wide association study (TWAS) using GTEx gene expression information. The colocalization analysis estimates the probability of shared signal between *cis*-

eQTL and GWAS analysis of testosterone(*37*). Using a probability of colocalization > 0.8, we identified 28 genes modulating testosterone levels (**Figure 3.4)**. Tissue types with most identified genes included the thyroid (8) and aorta (6). Additionally, genes were identified in organs relevant to testosterone synthesis, such as the testis (3), prostate (2), liver (2) and adrenal gland (2). The colocalization analysis revealed high colocalization between testosterone variants and *cis* variants for *SHBG, TNFSF12, BRI3, LAT, NRBF2,* and *BAIAP2L1*. *SHBG* variants have been strongly associated with variation in testosterone action(*38*, *39*). *BAIAP2L1* colocalization was specific to testes while *BRI3* colocalization was specific to the liver. Colocalization of *LAT* and *TNFSF12* was specific for the brain putamen/basal ganglia and brain spinal cord-cervical C1.

The TWAS implicated 214 genes and 22 tissue types in total testosterone regulation. Tissue types for which most TWAS associations were identified included skin (sun-exposed), thyroid and tibial artery, consistent with tissue types identified through colocalization analysis. Forty-six associations with Testis and 27 associations with the Liver were observed. Of 28 genes identified by colocalization analysis, 23 were also identified by TWAS analysis. *BRI3* and *BAIAP2L1,* which are both located at chromosome 7, were identified specifically through liver and testis TWAS analysis, specifically. *BAIAP2L1* (Brain-Specific Angiogenesis Inhibitor 1-Associated Protein 2-Like Protein 1) is expressed in urothelial and glandular cells and has been implicated in previous GWAS studies of testosterone(*40*). Gene expression studies identified *BAIAP2L1* to be upregulated with age and downregulated with riluzole treatment, a glutamate modulator shown to improve memory function in aged rats(*41*). *BRI3,* or Brain protein I3, is a negative regulator of amyloid precursor protein (APP)(*42*), providing a further link between the genetic determinants of testosterone regulation and Alzheimer's disease pathophysiology. Although these genes have been

implicated through tissues such as liver and testes, they might have broad effects on the human body and oncogenic mechanisms(*43–45*).

To determine ancestry-associated differences in genetic regulation of testosterone, we conducted a Network Assisted Genomic Association (NAGA) on the GWAS results from each ancestry group and compared the top 50 implicated genes. Twenty-five genes were shared across all groups, including *SHBG, TNFSF12, SENP3, JMJD1C, RRAG1,* and *TP53.* Unique genes for the African ancestry group included *MSN, DMD, GAGE2B and GAGE2C,* while unique genes for the Hispanic ancestry group included *SERPINA7, TKTL1, TTC28,* and *PNPLA3.* Of the genes uniquely implicated, 9 and 2 were on chromosome X for African and Hispanic ancestry groups, respectively.

3.1.3.4 PheWAS and LabWAS characterization of Testosterone Variants

Given the availability of extensive electronic health record data in the MVP, we conducted a phenome-wide association study (PheWAS) among individual level genotypes and ICD-10 code data to determine whether the 46 trans-ancestry testosterone variants were associated with disease risk . We identified 920, 45 and 56 significant associations (FDR < .01) in the European, African and Hispanic ancestry groups, respectively. Nine variants had PheWAS associations across all ancestry groups; 2 had associations uniquely in the African ancestry group. Across all ancestry groups, PheWAS associations for testicular dysfunction/hypofunction, hyperglyceridemia, gout, type 2 diabetes, chronic liver disease, hepatomegaly and hyperlipidemia were observed (**Figure 3.6)**. In general, testosterone-increasing variants across ancestry groups were associated with lower odds of testicular hypofunction, dysfunction, morbid obesity, hyperglyceridemia, sleep apnea, diabetes mellitus, gout, hyperlipidemia and increased odds of hepatomegaly, liver disease and

esophageal bleeding. Opposite to general trends, two variants on chromosome 2 and 17 (rs1260326, rs188272638) in *GCKR* and *POLR2A*, respectively, in the European ancestry group were associated with increased testosterone and increased obesity. Trans-ancestry variant rs190119169, an *AR* variant on chromosome X, was associated with increased testosterone across all ancestry groups (beta_eur=4.85, beta_afr=7.87, beta_his=6.11), but only significantly associated with lower odds of diabetes in the African ancestry group **(Figure 3.6)**. Two testosterone-increasing variants, rs112635299 on chromosome 14 and rs764029425 on chromosome 7 near the *BAIAP2L1* gene, were significantly associated with lower odds of prostate cancer in the European ancestry group (rs112635299 OR=0.89 [0.86-0.91], rs764029425 OR=0.95 [0.94-0.96]).

Since testosterone can influence lab measurements(*46*), we also ran LabWAS associations with 46 *trans*-ancestry variants and mean levels of lipid, metabolic and blood lab values. We identified 668 significant associations in the European, 150 in the African and 44 in the Hispanic ancestry groups (FDR<.01). Across all ancestry groups, significant associations with calcium, aspartate aminotransferase (AST), bicarbonate, A1c, hematocrit, estimated glomerular filtration rate (eGFR), alanine aminotransferase (ALT), HDL-C, iron, glucose, albumin, hemoglobin, LDL-C, international normalized ratio (INR), creatinine and chloride were observed. Of these shared associations, most associations with HDL-C, albumin and LDL-C were observed in the European ancestry group compared to glucose, hemoglobin and LDL-C in the African ancestry group. rs190119169, which was significantly associated with lower odds of diabetes in the African ancestry group, was also associated with significantly lower levels of mean glucose levels (beta=-5.14) in contrast to the European (beta=-0.60) and Hispanic (beta=-3.32) ancestry groups **(Figure 3.7)**.

We next evaluated PheWAS and LabWAS associations of 17 ancestry-specific variants. One of the 8 African-ancestry specific variants and 3 of the 5 Hispanic ancestry-specific variants had PheWAS association in respective ancestry groups. African-specific ancestry variant rs4827451 on chromosome X was uniquely associated with diabetes, diabetic retinopathy and renal failure in the African ancestry group. This variant was also uniquely associated with A1c, glucose and eGFR in the African ancestry group **(Figure 3.7)**. One of the genes in close proximity to rs4827451 is *MSN,* which could have potential mechanisms in renal fibrosis. rs73629199, a *TKTL1* variant on chromosome X, was uniquely associated with hereditary hemolytic anemias in the Hispanic ancestry group. Furthermore, significant associations were observed with hematocrit, ferritin, hemoglobin in the Hispanic ancestry group along with the African ancestry group.

3.1.3.5 Testosterone Polygenic Score (PGS) is Predictive of Testicular Dysfunction and Hyperlipidemia Across All Ancestry Groups

Given the numerous PheWAS and LabWAS associations with testosterone variants, we wanted to assess the predictive value of testosterone variants for different clinical phenotypes. We constructed a testosterone polygenic score from 46 trans-ancestry variants, which we validated in the MVP cohort **(Figure 3.7)** and the UK Biobank (**Figure S3.4).** Distribution of testosterone polygenic scores amongst ancestry groups were significantly different (t-test $p < 2e\text{-}16$). The African ancestry group had the highest average testosterone PGS scores (mean=7.76, std=0.26) followed by the European (mean=5.88,std=0.34) and Hispanic ancestry groups (mean=4.69, std=0.40) **(Figure S3.5)**.

To assess the predictive value of the testosterone PGS for clinical phenotypes, we conducted association analyses between the testosterone PGS and the clinical phenotypes with

PheWAS associations across ancestry group, namely testicular dysfunction, testicular hypofunction, hyperlipidemia, gout, Type 1 and 2 diabetes, and prostate cancer. The testosterone PGS was associated with all clinical phenotypes with the strongest effects observed for testicular dysfunction (OR_{eur} = 0.86 [0.85-0.87], OR_{afr} = 0.90 [0.87-0.92], OR_{his} = 0.84 [0.81-0.88]) and testicular hypofunction (OR_{eur}=0.85 [0.84-0.86], OR_{afr}=0.88 [0.85-0.90], OR_{his} = 0.84 [0.79-0.87]). Higher testosterone PGS was associated with lower odds of gout, hyperlipidemia, Type 1 and 2 diabetes, and prostate cancer **(Figure 3.7)**. Higher testosterone PGS was associated with significantly lower odds of gout among Hispanics as compared to the European and African ancestry groups. Furthermore, testosterone PGS was uniquely associated with lower odds of prostate cancer in the Hispanic and African ancestry groups, but not European ancestry group. These results suggest a differential role for testosterone regulation in disease risk stratification across ancestry groups.

3.1.4 Discussion

Utilizing data from the MVP cohort, we present the largest multi-ancestral genome-wide analysis of total testosterone levels to our knowledge. We identified 46 trans-ancestry variants of which 14 were novel. We identified 17 ancestry-specific variants in European (8), African (5), and Hispanic (4) ancestries of which 15 were novel. Of the 63 testosterone variants, 32 trans-ancestry and 2 ancestry-specific variants were in high LD (R2 > 0.2) with variants identified in UK Biobank analysis of testosterone levels(*19*). The genes implicated by shared variants include *SAT2, SHBG, FXR2, MPDU1, SOX15, TNFSF12-TNFSF13, SENP3, FGF11, CHRNB1, ZBTB4, POLR2A, TP53, JMJD1C, NRBF2, TDGF1P3, RGAG1,* and *AMMECR1. MSN, DMD, VSIG4,* and *HEPH*

were genes implicated by African ancestry group analysis specifically. *CHEK2, TTC28,* and *TKTL1* were genes implicated by Hispanic ancestry group analysis.

It is well-established that testosterone is immunosuppressive dampening many aspects of humoral- and cell-mediated immunity. Interestingly, both *VSIG4* and *MSN,* implicated uniquely by African ancestry group analysis, are immunomodulators. *VSIG4* is a macrophage-expressed complement receptor of the immunoglobulin superfamily with T-cell suppressive and differentiation effects(*47–49*) while *MSN* is an adaptor molecule essential for cell shape, motility and signaling regulation(*50–52*). *MSN* improves prognosis in lung adenocarcinoma patients by enhancing immune lymphocyte infiltration(*53*). Studies have linked testosterone with *MSN* expression, demonstrating the interaction promotes actin cytoskeletal remodeling and enhanced migration in endothelial cells(*54*). The immunomodulatory effects of testosterone have downstream effects for atherosclerosis, COVID-19 infection, and other diseases(*55*, *56*). Identification of immunomodulatory genes specifically in the African ancestry group suggests an ancestry-specific genetic link between testosterone and the immune system that is important to explore.

Our analysis found novel European ancestry-specific variants for the Ephrin B1 (EFNB1) gene at Xq13 and the WW Domain Containing Oxidoreductase (WWOX) gene at 16q23.1-2 that were associated with testosterone levels. The *EFNB1* gene encodes a ligand for Ephrin receptor tyrosine kinases, which has critical roles in migration and adhesion of brain neurons and cells in the lung, adipose tissue, kidney, and cardiovascular system during development. Recently, EFNB1 in hippocampal astrocytes has been found to regulate excitatory and inhibitory neurocircuits and exert a neuroprotective role in the adult brain(*57*, *58*). The WWOX protein contains a short–chain dehydrogenase reductase (SDR) domain at its C-terminus that is involved in androgen

steroidogenesis in Leydig cells of the testis; knockout of the WWOX results in a severe testosterone deficiency, impaired hematopoiesis and cholesterol metabolism, and renal failure(*59*). Furthermore, a recent genetic meta-analysis discovered that the WWOX locus has a genome-wide association with Alzheimer's disease(*60*). Low testosterone levels in men is a risk factor for dementia(*61*). Furthermore, testosterone replacement therapy (TRT) can modestly improve cognition(*62*), suggesting testosterone has a potential protective effect for dementia. Novel testosterone loci in genetic regions relevant to dementia may be due to the higher representation of disease in the MVP that are not available in other biobanks and highlight genetic links between testosterone and memory function.

The *SHBG* locus on chromosome 17 was significantly associated with testosterone levels across ancestry groups, however, differences in PheWAS associations at this locus were observed in the three ancestries. rs918121801 was only associated with testicular dysfunction and hypofunction in the African ancestry group while rs62059839, in close proximity to rs918121801, was associated with testicular dysfunction and hypofunction in the European ancestry group. High heterogeneity at the *SHBG* locus was calculated amongst ancestry groups and *SHBG* levels are known to be highly heritable(*63–65*). Thus, differences in PheWAS associations within the *SHBG* locus may reflect complex linkage disequilibrium patterns between ancestry groups(*66*) or a shared causal variant with different ancestry group specific markers. Causality analysis and fine mapping of the *SHBG* locus may help to better understand ancestry-specific differences in testosterone regulation.

We also linked ancestry-specific variants with differential disease risk in African and Hispanic ancestry groups. Specifically, chromosome X variant rs4827451 was associated with lower testosterone levels and eGFR and higher odds of chronic kidney disease and renal failure.

Hypogonadism is frequently observed in renal failure(*67*), increasing risk of anemia and cardiovascular disease. Furthermore, sex-specific and ancestry-specific disparities in CKD are important, with kidney function deteriorating faster in men compared to women and higher rates of renal failure occurring in African Americans compared to Caucasians. Pathophysiological mechanisms underlying renal failure are complex(*68*), but maintaining androgen regulation of renal function may have a therapeutic effect on chronic kidney disease improving medical management(*69*). Our identified variant suggests a potential genetic mechanism linking hypogonadism and chronic kidney disease in individuals of African ancestry group specifically. Furthermore, the variant is located in close proximity to *MSN,* which is involved in renal fibrosis through E-cadherin interaction(*70*). These findings demonstrate a potential ancestry-specific genetic mechanism in the kidney that requires further investigation.

A unique PheWAS association between a *TKTL1* variant and hereditary hemolytic anemias was identified in the Hispanic ancestry group. Lower testosterone levels are associated with higher rates of anemia in men, while testosterone replacement therapy stimulates iron-dependent erythropoiesis and has beneficial effects on anemia. Given the high heritability of red blood cell (RBC) traits(*71*), the identified *TKTL1* variant may give insight into androgen-related genetic regulation of anemia in non-European individuals. Furthermore, rs73629199 was associated with RBC-related lab values such as ferritin, hematocrit, and hemoglobin that links testosterone with RBC regulation.

Of the 17 ancestry-specific variants, 8 were identified in the African, 5 were identified in Hispanic, and 4 were identified in European ancestry group analyses. Most African ancestry-specific variants were found on chromosome X, however, we discovered *LIN01414* was a unique variant on chromosome 8. Based on LD score regression analysis, heritability of testosterone levels

were highest in the African ancestry group; however, significant variants only explain a minority of heritability (3.3%). Even using a suggestive threshold, heritability explained was 9.1% compared to 24.6% and 30.4% in European and Hispanic ancestry groups, respectively. Differences in total testosterone levels have been observed in African individuals and several of the diseases implicated by PheWAS analysis of testosterone levels, such as diabetes, cardiovascular disease, and chronic renal disease, occur at higher rates in individuals of African ancestry(*72–81*). Understanding androgen-related regulation of diseases in individuals of African ancestry is critical and will require larger sample sizes.

We constructed a polygenic score (PGS) for testosterone using 46 trans-ancestry variants and observed differential associations with gout and diabetes in the Hispanic ancestry group compared to European and African ancestry groups. Furthermore, PGS was associated with lower odds of prostate cancer in the Hispanic and African ancestry groups compared to the European ancestry group. In hyperlipidemia and testicular dysfunction, PRS associations were more concordant across ancestry groups. These findings highlight shared effects of testosterone on disease risk, but also reveal differences that give insight into ancestry group-specific genetic architecture of testosterone.

The Million Veteran Program provides a breadth of clinical and genomic data that has allowed us to conduct the most ancestrally diverse genetic analysis of total testosterone levels to date. We note limitations with exploring ancestry group-specific mechanisms as the tissue-specific gene expression data that are available come predominantly from individuals with European ancestry. Thus, our TWAS and colocalization analysis rely on European gene expression data; however, gene expression patterns may be influenced by ancestry group differences. Furthermore, our PheWAS analysis focused on ICD10 codes for clinical phenotyping. These clinical outcomes

are important for assessing disease risk; however, testosterone can also impact disease severity, treatment response and disease progression. Future directions should include continuing clinical phenotyping efforts in the MVP and assessing interactions with genomic determinants of testosterone and its impact of health and disease.

In summary, we conducted a large-scale multi-ancestry GWAS of morning total testosterone levels in men from the Million Veteran Program and discovered multiple novel trans-ancestry and ancestry-specific variants influencing testosterone levels. Using additional clinical and gene expression data, we linked ancestry-group germline variant differences with differential disease risk and constructed a polygenic risk score predictive of a variety of clinical phenotypes. These results provide unique insights into not only testosterone regulation but also ancestry group differences in genetic regulation of testosterone.

3.1.5 Materials and Methods

3.1.5.1 Genome-Wide Association Analysis

For Million Veteran Program (MVP) genome-wide association studies (GWAS), total testosterone levels (ng/dl) between 7AM and 12PM for 88,385 individuals of European ancestry, 25,235 individuals of African ancestry and 10,973 individuals of Hispanic ancestry as determined by HARE groups were inverse rank normalized. PLINK glm method was applied to conduct association analyses with testosterone levels excluding variants of minor allele frequency threshold less than 0.1% and correcting for covariates: principal components (1-10), age. Patients with X chromosome aneuploidy (XXY, XYY) were excluded from analysis. Patients on androgen deprivation therapy (ADT), and testosterone replacement therapy (TRT) were excluded from

analysis to focus on analysis of endogenous testosterone levels. In patients with multiple testosterone levels readings, first testosterone reading was used.

For European ancestral group testosterone loci validation, field ID 30850 (Testosterone nmol/L) was extracted and inverse-rank normalized from UK Biobank. PLINK(*82*) glm method was used to calculate beta and p-values for variants identified in MVP discovery analysis. UK Biobank validation cohort consisted of 166,502 European, unrelated men. Top 10 principal components (PC1-PC10) and age were used as covariates.

For African ancestral group testosterone loci validation, total testosterone values were extracted and inverse-rank normalized for self-identifying African individuals in the Multi-Ethnic Study of Atherosclerosis (MESA). Individuals with KING relatedness > 0.177 were removed. PLINK glm method was used to conduct validation analysis of testosterone loci in 1317 men. Age was used as a covariate.

3.1.5.2 Conditional X chromosome analysis

PLINKv2.0 glm method was used to run stepwise conditional analysis for identification of independent X chromosome associations(*83*). First, association with testosterone was conducted and SNP with lowest p-value was identified. Next association with testosterone was run conditioning on most significant variants. Analysis was conducted until no SNPs with p-value < $1x10^{-6}$ remained. To reduce the chance of multicollinearity, variants which were within 50kb of the most recent significant variant were excluded from the selection process.

3.1.5.3 LDSC

LD scores for European, African and Hispanic ancestral groups were calculated using LDSC(*30*) from 1000 Genomes EUR, AFR, and AMR groups, respectively. For partitioned heritability analysis, categories for analysis included base and significant category for variants which based p-value threshold ($p < 5e-08$) for each ancestral group(*84*). Genetic correlation analysis was conducted with summary statistics for each ancestral group using LD scores from both groups in analysis(*35*). Lastly, for LD hub analysis, summary statistics were uploaded to LD Hub server (http://ldsc.broadinstitute.org/). Genetic correlations in UK Biobanks and "reproductive" category were excluded. FDR was calculated from LD hub provided p-values.

3.1.5.4 Colocalization Analysis

Significant variants ($p < 5e-08$) for each ancestral group were compared with significant GTEx eQTL results (accession date: 10/02/2019). All genes for which significant eQTL association was shared were included in colocalization analysis. Full summary statistics for GTEx *cis*-eQTL analysis were used to make input files including samples sizes and p-values for both GTEx and MVP testosterone summary statistics from European ancestral group. Minor allele frequency from GTEx consortium was included for final coloc analysis. coloc.abf function in coloc R package(*37*) was used to conduct analysis. Genes with PP.H4 (posterior probability of one common causal variant between studies) of 0.8 were analyzed.

3.1.5.5 TWAS

TWAS R package was downloaded from TWAS hub using 1000 genomes PLINK files to infer linkage disequilibrium patterns and 22 GTEx expression models(*85*). TWAS analysis was

conducted with all expression models using European linkage-disequilibrium reference from 1000 Genomes.

3.1.5.6 PheWAS & LabWAS

1817 International Classification of Diagnoses-10-Clinical Modification (ICD-10 CM) diagnosis codes from electronic health care records were available for MVP participants from as early as 1998. PheWAS R package was used to conduct PheWAS associations in male MVP European (n=421,212), African (n=104,380), Hispanic (n=45,553) ancestral groups. Cases were individuals with investigated phecode, whereas controls did not have incidence of phecode. Covariates included age at testosterone lab and first 10 genetic principal components.

Mean lab values for 68 measurements were available. As was done for the PheWAS, associations with mean lab values were conducted with PLINK using top 10 principal components and age as covariates.

False discovery rates (FDR) for PheWAS and LabWAS associations were calculated using Benjamini-Hochberg correction(*86*) with python statsmodel package(*87*). Associations with FDR less than 1% were considered significant.

3.1.5.7 Testosterone Polygenic Score

Meta-analysis of European, African and Hispanic testosterone GWAS were conducted with METAL. Variants passing significance threshold ($p<5e-08$) in meta-analysis were used to construct polygenic scores. 46 trans-ancestry variants were extracted from imputed genotype BGEN files. Alleles were oriented to increasing testosterone levels. Genotype dosages were weighted by beta values for each respective ancestry group and summed. Logistic regression

analyses with phecodes extracted from PheWAS dataframe were conducted with age of enrollment and top 10 principal components as covariates.

3.1.6 Figures

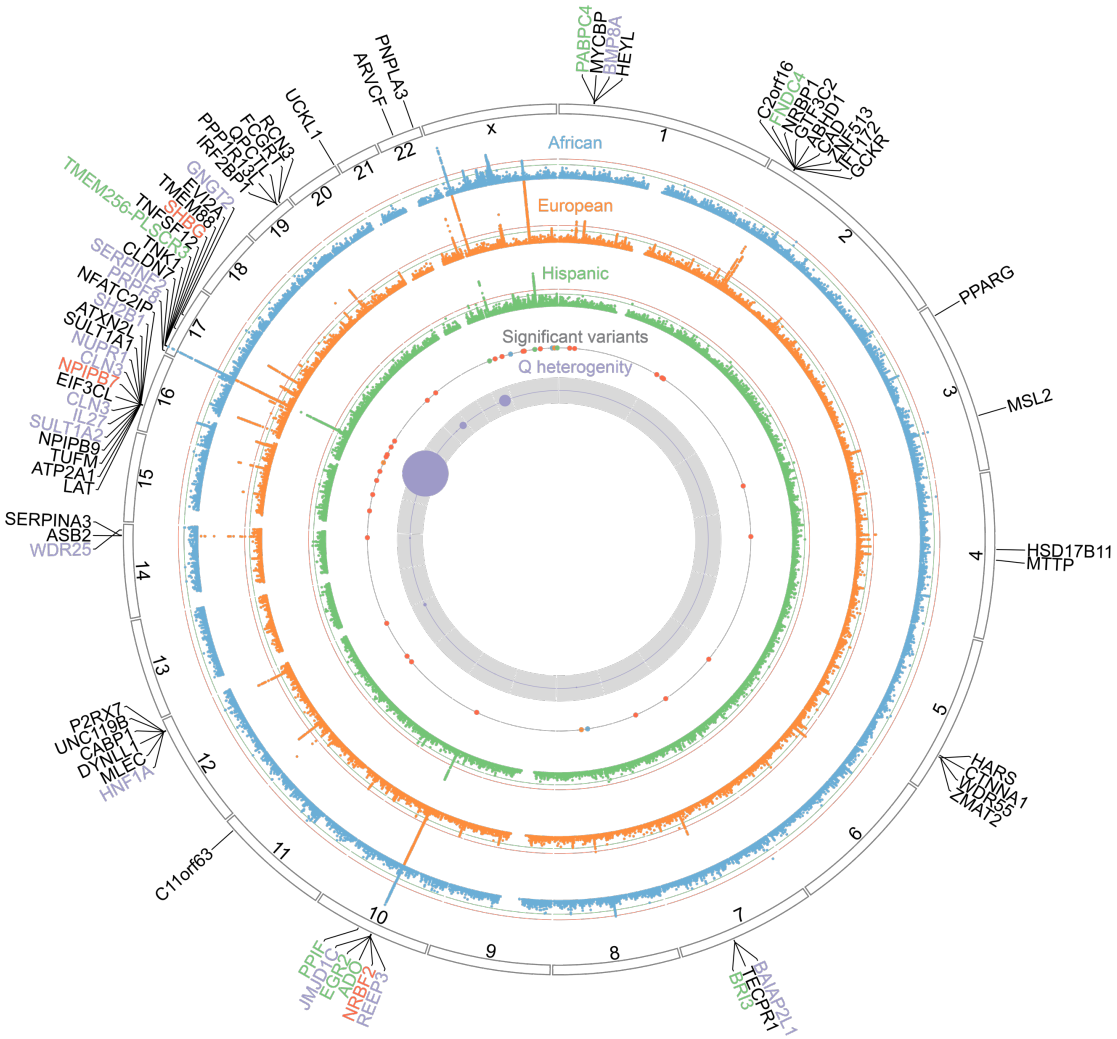

Figure 3.1 GWAS of Total Testosterone Levels in European, African and Hispanic HARE Groups Circos plot of -log10 p values of Total Testosterone GWAS in European (orange), African (blue) and Hispanic (green) ancestry groups in the Million Veteran Project (MVP). The green line indicates the suggestive threshold (1e-06) and the red line indicates the genome-wide significance threshold (5e-08). The innermost purple band corresponds to measures of heterogeneity between ancestry group GWAS using Cochran's Q test. Second innermost ring represents significant variants: red represents trans-ancestry variants, green represents Hispanic-specific variants, blue represents African-specific variants and orange represents European-specific variants. Selected genes near testosterone loci are indicated on the outside edge. A complete list of associations and nearby genes can be found in Supplemental Table 1.

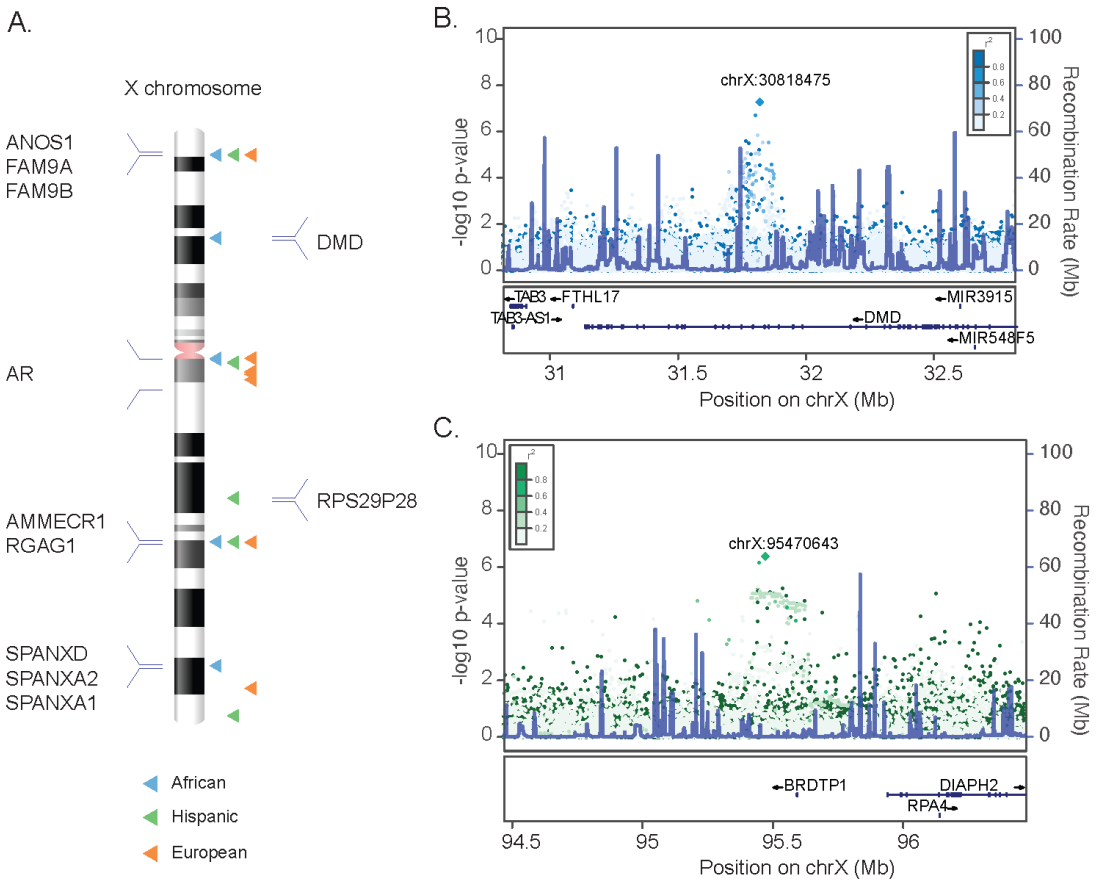

Figure 3.2 Sex Chromosome Analyses of Total Testosterone in European, African and Hispanic HARE Groups (A) Karyotype plot of conditional X chromosome variants identified in European, African and Hispanic HARE groups. **(B)** Locuszoom plot of a conditional X chromosome variant in *DMD* uniquely identified in the African ancestry group. **(C)** Locuszoom plot of a conditional X chromosome variant near the *RPS29P28* pseudogene uniquely identified in the Hispanic ancestry group. A complete list of conditional associations can be found in Supplemental Table 2.

A.

Group1	Group2	LDscore Reference	Correlation	P
European	African	European	0.19	0.0063
African	Hispanic	African	0.37	0.0809
European	Hispanic	European	0.6	0.0007

B.

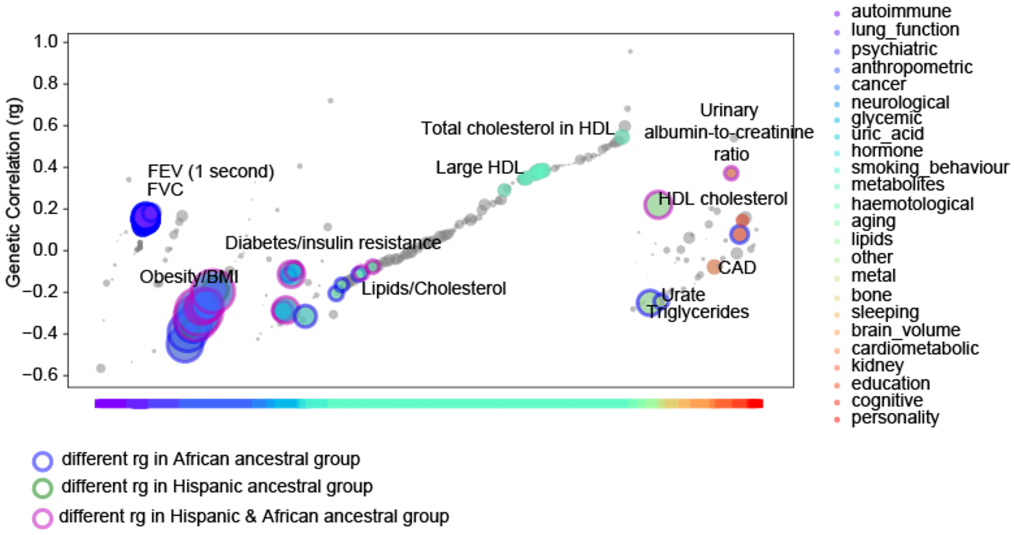

Figure 3.3 Genetic Correlation Across Ancestry Groups (A) Genetic correlation between ancestry groups estimated by LSDC **(B)** Genetic correlation between testosterone and other traits from LDhub[36] for European HARE groups. Genetic correlations different between ancestry groups are outlined in blue, if in African ancestry group only, green, if in Hispanic ancestry group only, and magenta is both in African and Hispanic ancestry groups. Selected categories are labels and full LDHub results are in Supplementary Tables 3-5.

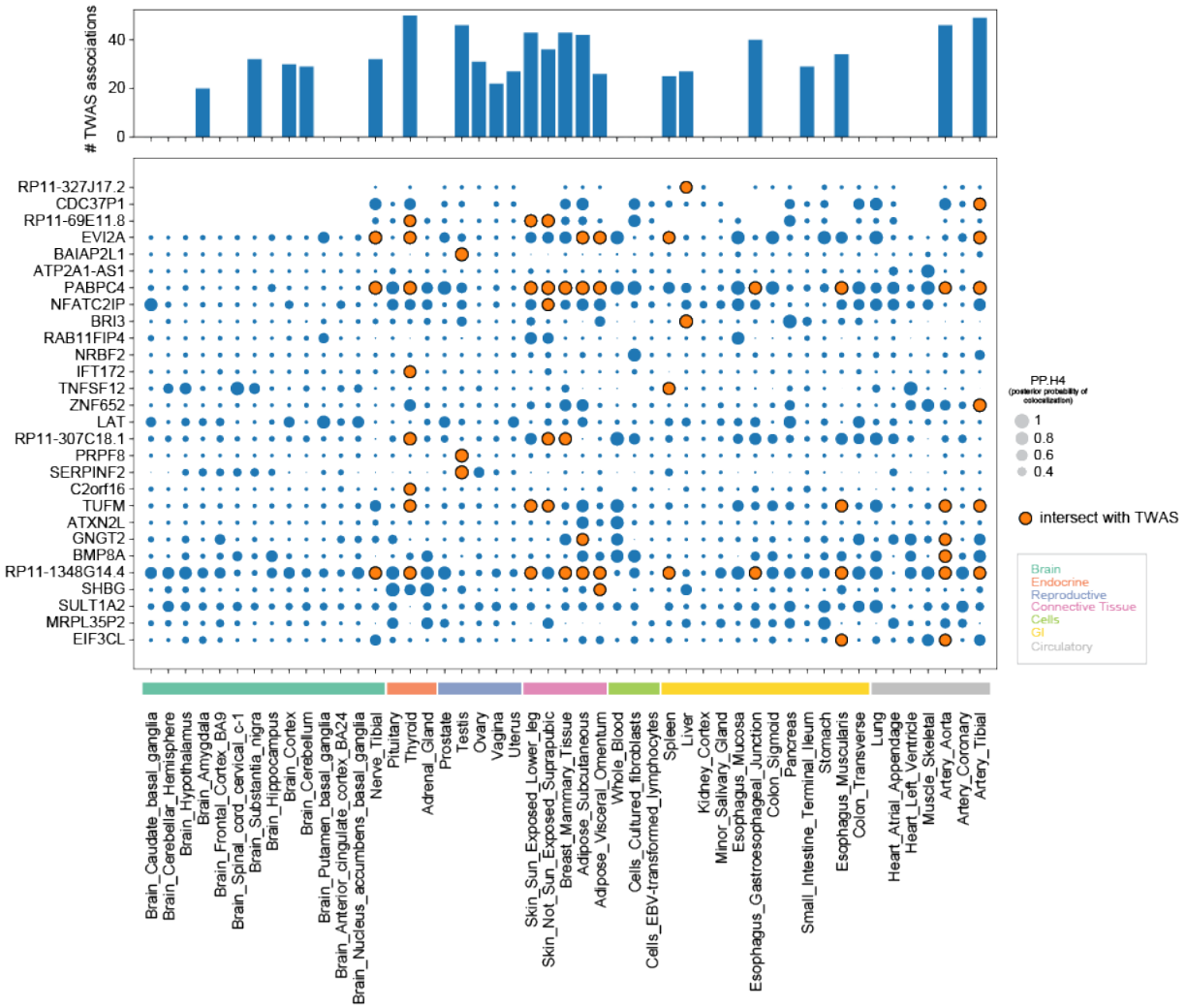

Figure 3.4 Testosterone Variants Affect Gene Expression in Specific Cell Types Grid plot of genes with high colocalization (PP.H4 [posterior probability both traits are associated and share a single causal variant]>0.8) between Testosterone and GTEx *cis*-eQTL association by tissue type. Points in orange were also implicated by TWAS. Top barplot describes the number of TWAS associations for each cell type. Full colocalization and TWAS results are available in Supplemental Table 6 and 7, respectively.

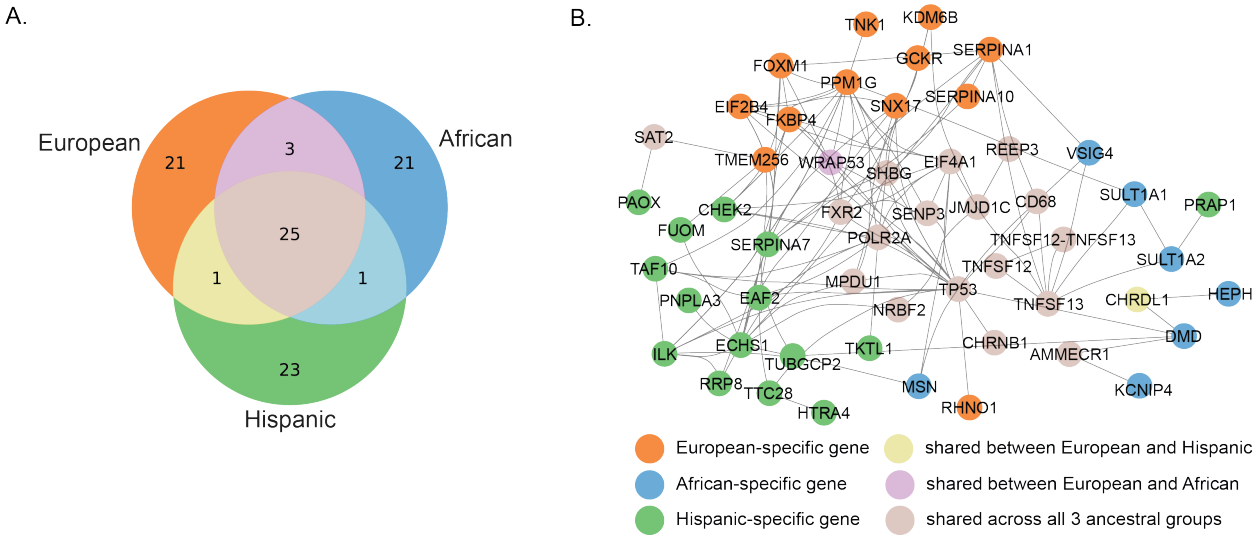

Figure 3.5 Shared Testosterone Genes Implicated by Network-Assisted Genomic Association (NAGA) (A) Venn diagram of top 50 genes implicated by NAGA analysis. **(B)** Fully connected subnetwork of genes implicated by NAGA. Nodes colored purple if shared across all 3 ancestry groups. European, African and Hispanic ancestry group-specific associations are colored orange, blue and green, respectively.

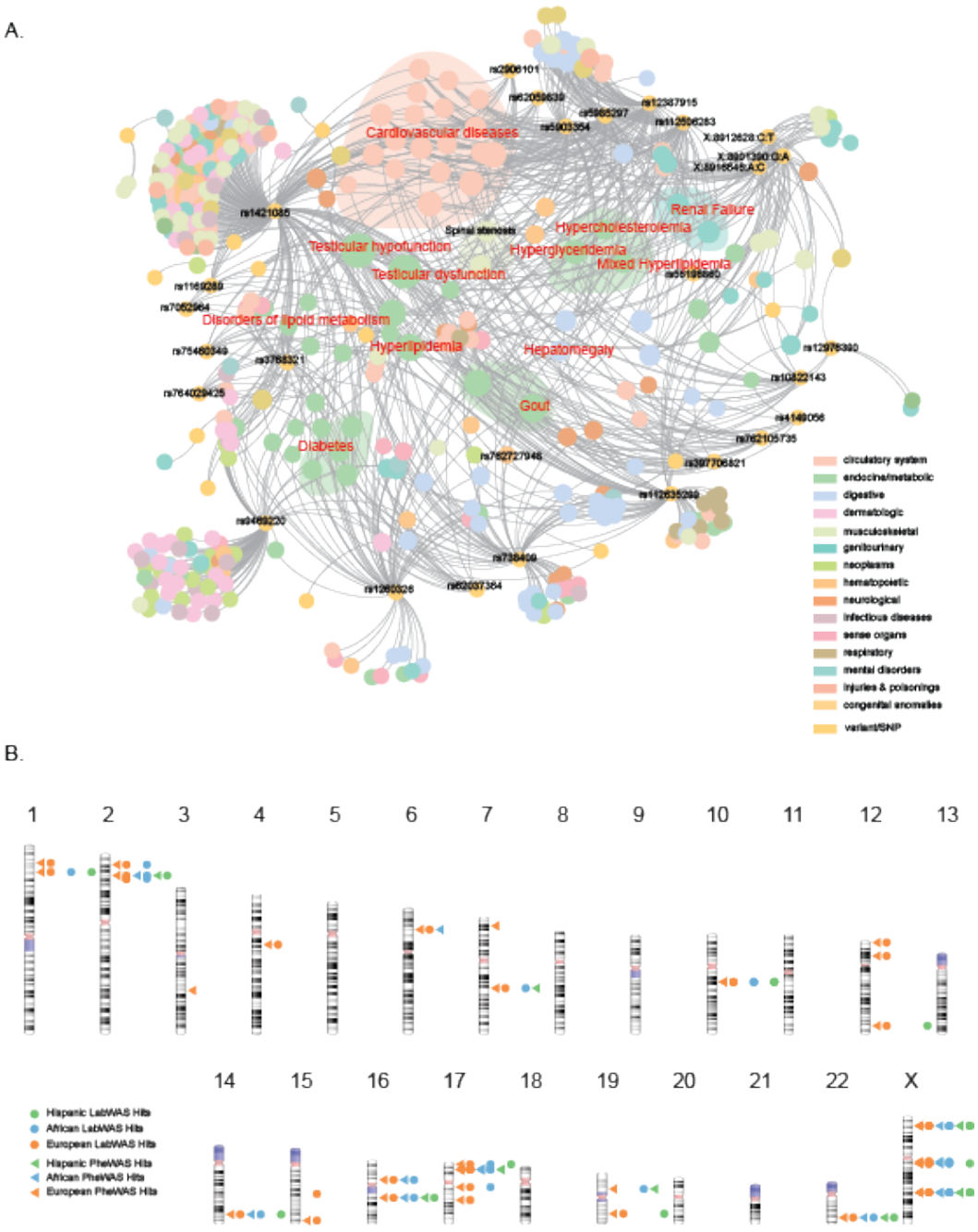

Figure 3.6 PheWAS Associations Highlight Major Diseases where Testosterone Genetic Variants are Implicated (A) Gephi network plot of PheWAS associations in European ancestry group. **(B)** Karyotype plot of testosterone variants (red highlight) and any PheWAS association (triangles/circles). Full PheWAS and LabWAS results are found in Supplemental Tables 9-14.

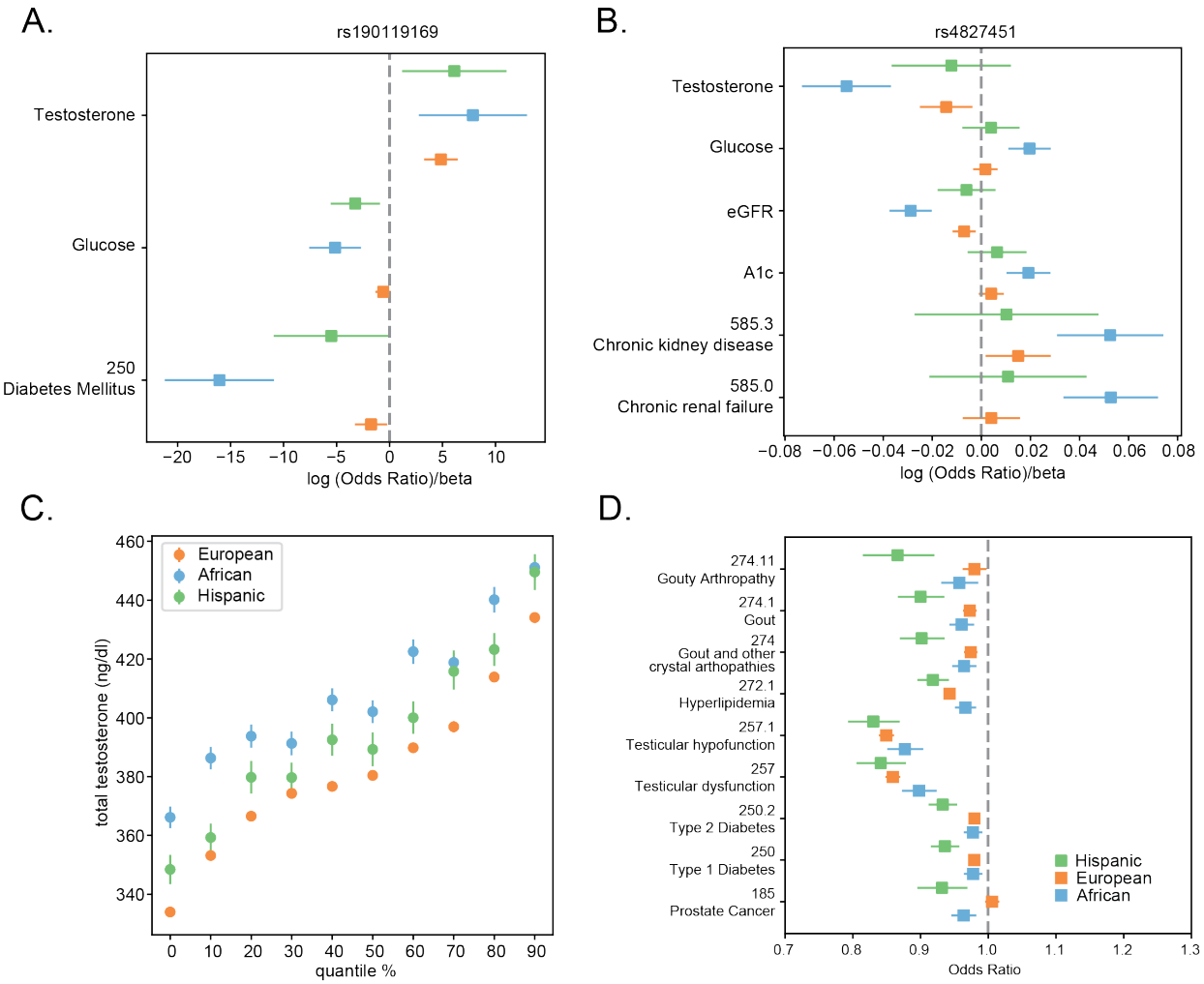

Figure 3.7 Ancestry-Specific Effects of Testosterone Variants and Polygenic Score (PGS) (A) Plot of chromosome X variant rs190119169 ancestry-specific effects on testosterone levels, average glucose lab levels and diabetes risk. **(B)** Plot of chromosome X variant rs4827451 ancestry-specific effects on testosterone levels, average glucose levels, average eGFR, A1c levels and chronic kidney disease and renal failure risk. **(C)** Quantile plot of testosterone PGS and testosterone levels (ng/dl) in the MVP discovery cohort. **(D)** Odds ratio plot of the testosterone PGS association with testicular dysfunction, testicular hypofunction, hyperlipidemia, gout, type 1 and 2 diabetes, and prostate cancer.

3.1.7 Tables

Table 3.1 MVP Participant Characteristics for Testosterone GWAS Table of MVP participants in European, African and Hispanic HARE groups used for testosterone GWAS.

	European	**African**	**Hispanic**
# male participants	426,223	104,882	45,890
# male participants with AM total testosterone levels	88,385	25,235	10,973
Mean total testosterone levels (Standard Deviation) ng/dl	383 (189)	408 (208)	394 (189)
Mean age of testosterone lab (Standard Deviation) years	60.45 (11.55)	55.82 (9.93)	54.72 (12.68)

Table 3.2 Partitioned Heritability Results from LDSC for European, African and Hispanic HARE Cohorts

		h2 explained	
Ancestral group	**H2 (heritability)**	**significant snps (5e-08)**	**suggestive snps (1e-05)**
European (n=88,385)	7.08 (0.71)%	14.6 (7.6)%	24.6 (6.4)%
African (n=25,235)	9.48 (3.58)%	3.3 (16.3)%	9.1 (18.0)%
Hispanic (n=10,973)	8.40 (2.89)%	11.1 (7.3)%	30.4 (10.6)%

3.1.8 Supplementary Figures

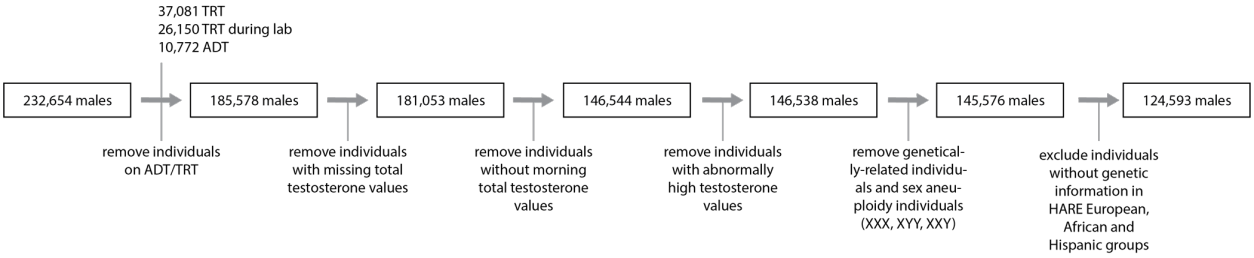

Figure S3.1 Workflow of Testosterone Phenotype Processing in Million Veteran Program Steps to process testosterone information from Million Veteran Program. Individuals on androgen deprivation therapy (ADT), testosterone replacement therapy (TRT) and TRT during testosterone lab measurement were excluded. Analysis was focused on morning (7 AM – 12 PM) testosterone levels. Related individuals and individuals with sex aneuploidies were excluded.

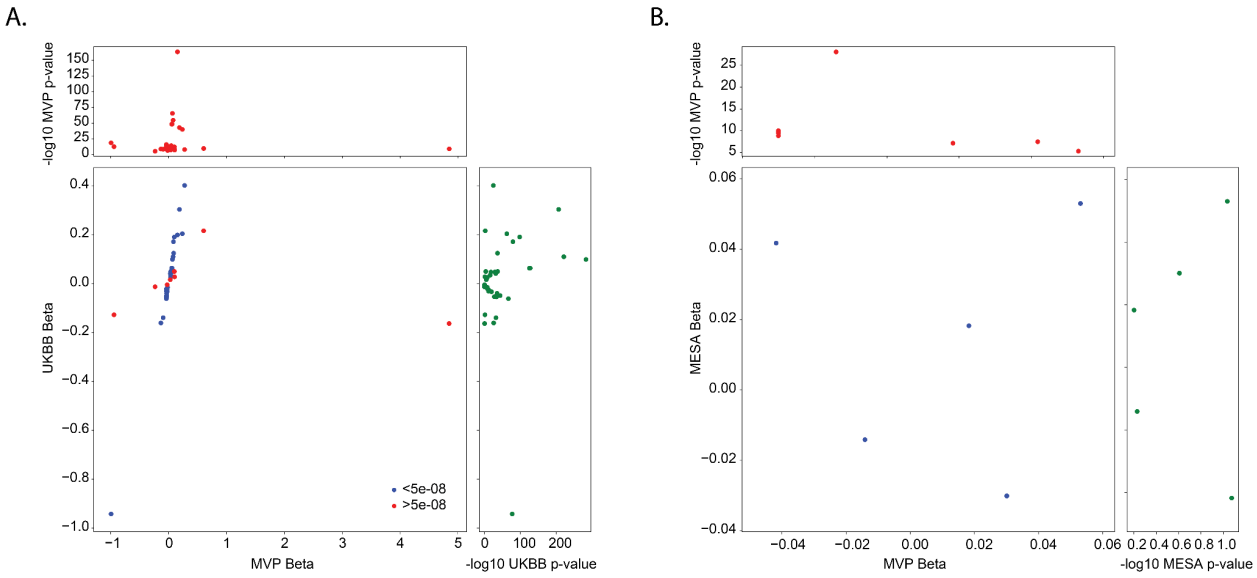

Figure S3.2 Validation of Testosterone GWAS Hits in UKBB and MESA cohort (A) Scatterplot of MVP Testosterone beta values and UKBB Testosterone beta values from GWAS conducted in only males. Points in blue pass significance threshold (5e-08) in UK Biobank. **(B).** Scatterplot of MVP Testosterone beta values and MESA cohort Testosterone beta values from GWAS conducted in only African males. No points pass significance threshold (5e-08).

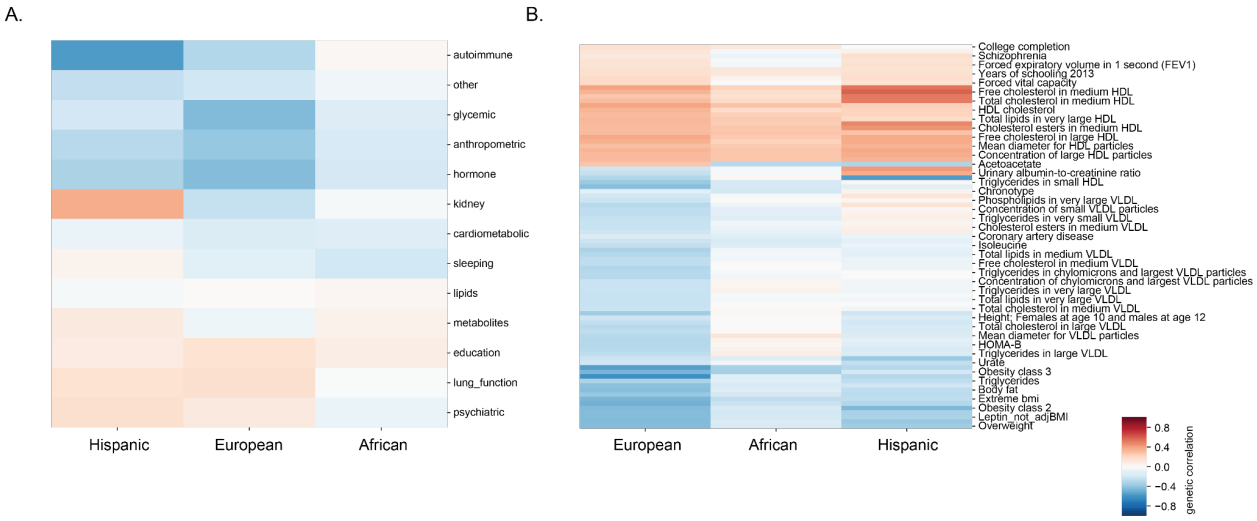

Figure S3.3 LDhub Correlations Amongst European, African and Hispanic Ancestral Groups (A) Heatmap of genetic correlations of GWAS catalog categories. (B) Heatmap of genetic correlations of GWAS catalogs.

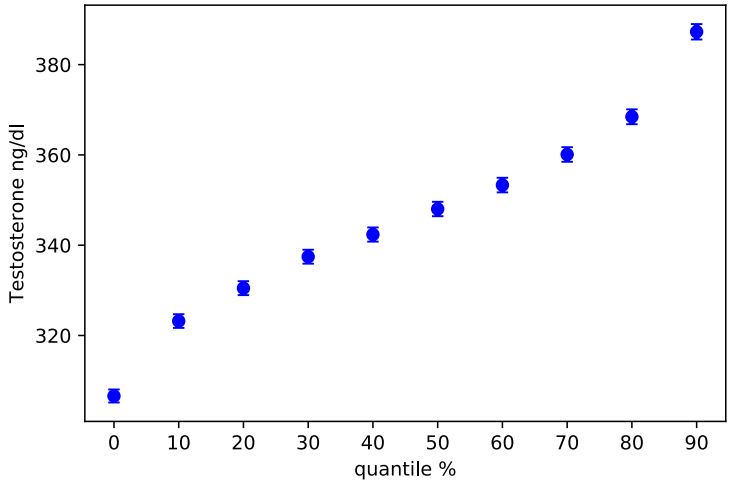

Figure S3.4 Testosterone Polygenic Score (PGS) Validation in UK Biobank Quantile plot of testosterone PGS constructed from 37 of the 46 trans-ancestry SNPs and testosterone levels in UK Biobank.

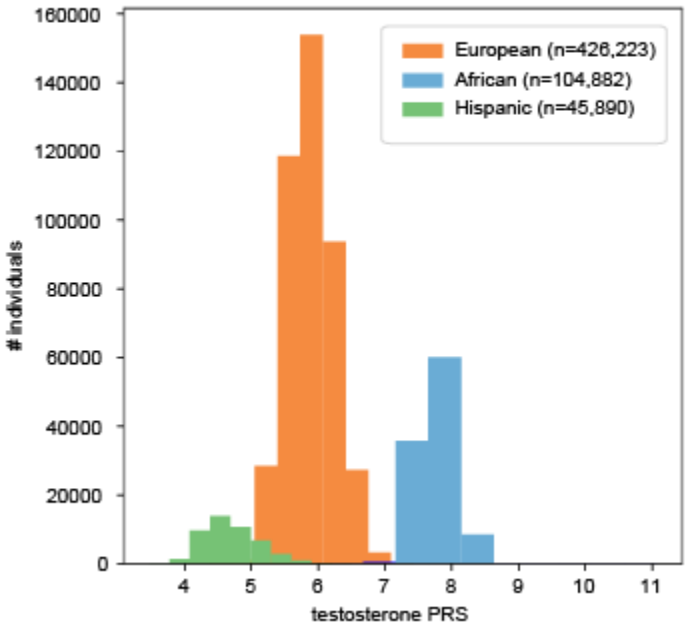

Figure S3.5 Testosterone Polygenic Score (PGS) Distribution in European, African and Hispanic Ancestry Groups of Million Veteran Program (MVP) Histogram of testosterone PGS constructed from 46 variants in European, African and Hispanic ancestry groups in Million Veteran Program (MVP).

3.1.9 Author Contributions

Original concept: Meghana Pagadala, Richard Hauger, Matthew Panizzon

Clinical Phenotyping: Nelson Chang, Rishi Deka, Kyung Lee, Tori Anglin, Julie Lynch

Genetic Analysis: Meghana Pagadala

Manuscript writing: Guneet Kaur, Tyler Seibert, Brent Rose, Julie Lynch, Hannah Carter, Matthew Panizzon, Richard Hauger

3.1.10 Acknowledgements

This research used data from the Million Veteran Program, Office of Research and Development, Veterans Health Administration. This research was supported by the Million Veteran Program MVP022 award # I01 CX001727 (PI: Richard L. Hauger MD). This publication does not represent the views of the Department of Veterans Affairs or the United States Government. Dr. Hauger was additionally funded by the VISN-22 VA Center of Excellence for Stress and Mental Health (CESAMH) and National Institute of Aging RO1 grant AG050595 (*The VETSA Longitudinal Twin Study of Cognition and Aging VETSA 4).*

For LDhub analysis, we gratefully acknowledge all the studies and databases that made GWAS summary data available: ADIPOGen (Adiponectin genetics consortium), C4D (Coronary Artery Disease Genetics Consortium), CARDIoGRAM (Coronary ARtery DIsease Genome wide Replication and Meta-analysis), CKDGen (Chronic Kidney Disease Genetics consortium), dbGAP (database of Genotypes and Phenotypes), DIAGRAM (DIAbetes Genetics Replication And Meta-analysis), ENIGMA (Enhancing Neuro Imaging Genetics through Meta Analysis), EAGLE (EArly Genetics & Lifecourse Epidemiology Eczema Consortium, excluding 23andMe), EGG (Early Growth Genetics Consortium), GABRIEL (A Multidisciplinary Study to Identify the Genetic and Environmental Causes of Asthma in the European Community), GCAN (Genetic Consortium for Anorexia Nervosa), GEFOS (GEnetic Factors for OSteoporosis Consortium), GIANT (Genetic Investigation of ANthropometric Traits), GIS (Genetics of Iron Status consortium), GLGC (Global Lipids Genetics Consortium), GPC (Genetics of Personality Consortium), GUGC (Global Urate and Gout consortium), HaemGen (haemotological and platelet traits genetics consortium), HRgene (Heart Rate consortium), IIBDGC (International Inflammatory Bowel Disease Genetics

Consortium), ILCCO (International Lung Cancer Consortium), IMSGC (International Multiple Sclerosis Genetic Consortium), MAGIC (Meta-Analyses of Glucose and Insulin-related traits Consortium), MESA (Multi-Ethnic Study of Atherosclerosis), PGC (Psychiatric Genomics Consortium), Project MinE consortium, ReproGen (Reproductive Genetics Consortium), SSGAC (Social Science Genetics Association Consortium) and TAG (Tobacco and Genetics Consortium), TRICL (Transdisciplinary Research in Cancer of the Lung consortium), UK Biobank. We gratefully acknowledge the contributions of Alkes Price (the systemic lupus erythematosus GWAS and primary biliary cirrhosis GWAS) and Johannes Kettunen (lipids metabolites GWAS).

The content of this chapter, in full, is a reprint of the this publication as it appears in medrxiv: Pagadala, Meghana; Jasuja, Guneet; Palnati, Madhuri; Lynch, Julie; Anglin, Tori; Chang, Nelson; Deka, Rishi; Lee, Kyung; Agiri, Fati; Seibert, Tyler; Rose, Brent; Carter, Hannah; Panizzon, Matthew; Hauger, Richard. Discovery of Novel Trans-Ancestry and Ancestry-Specific Gene Loci for Total Testosterone in a Multi-Ancestral Analysis of Men in the Million Veteran Program.

3.1.11 References

1. S. Bhasin, J. P. Brito, G. R. Cunningham, F. J. Hayes, H. N. Hodis, A. M. Matsumoto, P. J. Snyder, R. S. Swerdloff, F. C. Wu, M. A. Yialamas, Testosterone Therapy in Men With Hypogonadism: An Endocrine Society Clinical Practice Guideline. *J. Clin. Endocrinol. Metab.* **103**, 1715–1744 (2018).

2. A. M. Matsumoto, W. J. Bremner, Testicular Disorders. *Williams Textbook of Endocrinology* (2011), pp. 688–777.

3. J. D. Wilson, J. E. Griffin, F. WGeorge, M. Leshin, The Endocrine Control of Male Phenotypic Development. *Australian Journal of Biological Sciences*. **36** (1983), p. 101.

4. S. Shahani, M. Braga-Basaria, M. Maggio, S. Basaria, Androgens and erythropoiesis: past and present. *J. Endocrinol. Invest.* **32**, 704–716 (2009).

5. K. L. Herbst, S. Bhasin, Testosterone action on skeletal muscle. *Current Opinion in Clinical Nutrition and Metabolic Care*. **7** (2004), pp. 271–277.

6. M. R. Gubbels Bupp, T. N. Jorgensen, Androgen-Induced Immunosuppression. *Front. Immunol.* **9**, 794 (2018).

7. D. M. Kelly, T. H. Jones, Testosterone: a vascular hormone in health and disease. *J. Endocrinol.* **217**, R47–71 (2013).

8. M. W. O'Reilly, P. J. House, J. W. Tomlinson, Understanding androgen action in adipose tissue. *J. Steroid Biochem. Mol. Biol.* **143**, 277–284 (2014).

9. R. A. Davey, M. Grossmann, Androgen Receptor Structure, Function and Biology: From Bench to Bedside. *Clin. Biochem. Rev.* **37**, 3–15 (2016).

10. J. M. Kaufman, A. Vermeulen, The decline of androgen levels in elderly men and its clinical and therapeutic implications. *Endocr. Rev.* **26**, 833–876 (2005).

11. J.-M. Kaufman, B. Lapauw, A. Mahmoud, G. T'Sjoen, I. T. Huhtaniemi, Aging and the Male Reproductive System. *Endocr. Rev.* **40**, 906–972 (2019).

12. I. Huhtaniemi, G. Forti, Male late-onset hypogonadism: pathogenesis, diagnosis and treatment. *Nature Reviews Urology*. **8** (2011), pp. 335–344.

13. A. Morgentaler, E. L. Rhoden, Prevalence of prostate cancer among hypogonadal men with prostate-specific antigen levels of 4.0 ng/mL or less. *Urology*. **68**, 1263–1267 (2006).

14. K. T. Nead, Androgens and depression: a review and update. *Curr. Opin. Endocrinol. Diabetes Obes.* **26**, 175–179 (2019).

15. M. S. Panizzon, R. Hauger, A. M. Dale, L. J. Eaves, L. T. Eyler, B. Fischl, C. Fennema-

Notestine, C. E. Franz, M. D. Grant, A. J. Jak, K. C. Jacobson, M. J. Lyons, S. P. Mendoza, M. C. Neale, E. C. Prom-Wormley, L. J. Seidman, M. T. Tsuang, H. Xian, W. S. Kremen, Testosterone modifies the effect of APOE genotype on hippocampal volume in middle-aged men. *Neurology*. **75**, 874–880 (2010).

16. E. Vegeto, A. Villa, S. Della Torre, V. Crippa, P. Rusmini, R. Cristofani, M. Galbiati, A. Maggi, A. Poletti, The Role of Sex and Sex Hormones in Neurodegenerative Diseases. *Endocr. Rev.* **41** (2020), doi:10.1210/endrev/bnz005.

17. R. S. Vest, C. J. Pike, Gender, sex steroid hormones, and Alzheimer's disease. *Hormones and Behavior*. **63** (2013), pp. 301–307.

18. B. B. Yeap, Hormonal changes and their impact on cognition and mental health of ageing men. *Maturitas*. **79**, 227–235 (2014).

19. K. S. Ruth, F. R. Day, J. Tyrrell, D. J. Thompson, A. R. Wood, A. Mahajan, R. N. Beaumont, L. Wittemans, S. Martin, A. S. Busch, A. M. Erzurumluoglu, B. Hollis, T. A. O'Mara, Endometrial Cancer Association Consortium, M. I. McCarthy, C. Langenberg, D. F. Easton, N. J. Wareham, S. Burgess, A. Murray, K. K. Ong, T. M. Frayling, J. R. B. Perry, Using human genetics to understand the disease impacts of testosterone in men and women. *Nat. Med.* **26**, 252–258 (2020).

20. N. Sinnott-Armstrong, S. Naqvi, M. Rivas, J. K. Pritchard, GWAS of three molecular traits highlights core genes and pathways alongside a highly polygenic background. *Elife*. **10** (2021), doi:10.7554/eLife.58615.

21. B. P. Prins, K. B. Kuchenbaecker, Y. Bao, M. Smart, D. Zabaneh, G. Fatemifar, J. 'an Luan, N. J. Wareham, R. A. Scott, J. R. B. Perry, C. Langenberg, M. Benzeval, M. Kumari, E. Zeggini, Genome-wide analysis of health-related biomarkers in the UK Household Longitudinal Study reveals novel associations. *Sci. Rep.* **7**, 11008 (2017).

22. E. Flynn, Y. Tanigawa, F. Rodriguez, R. B. Altman, N. Sinnott-Armstrong, M. A. Rivas, Sex-specific genetic effects across biomarkers. *Eur. J. Hum. Genet.* **29**, 154–163 (2021).

23. G. Jin, J. Sun, S.-T. Kim, J. Feng, Z. Wang, S. Tao, Z. Chen, L. Purcell, S. Smith, W. B. Isaacs, R. S. Rittmaster, S. L. Zheng, L. D. Condreay, J. Xu, Genome-wide association study identifies a new locus JMJD1C at 10q21 that may influence serum androgen levels in men. *Hum. Mol. Genet.* **21**, 5222–5228 (2012).

24. C. Ohlsson, H. Wallaschofski, K. L. Lunetta, L. Stolk, J. R. B. Perry, A. Koster, A.-K. Petersen, J. Eriksson, T. Lehtimäki, I. T. Huhtaniemi, G. L. Hammond, M. Maggio, A. D. Coviello, EMAS Study Group, L. Ferrucci, M. Heier, A. Hofman, K. L. Holliday, J.-O. Jansson, M. Kähönen, D. Karasik, M. K. Karlsson, D. P. Kiel, Y. Liu, O. Ljunggren, M. Lorentzon, L.-P. Lyytikäinen, T. Meitinger, D. Mellström, D. Melzer, I. Miljkovic, M. Nauck, M. Nilsson, B. Penninx, S. R. Pye, R. S. Vasan, M. Reincke, F. Rivadeneira, A. Tajar, A. Teumer, A. G. Uitterlinden, J. Ulloor, J. Viikari, U. Völker, H. Völzke, H. E. Wichmann, T.-S. Wu, W. V. Zhuang, E. Ziv, F. C. W. Wu, O. Raitakari, A. Eriksson, M. Bidlingmaier, T. B. Harris, A. Murray, F. H. de Jong, J. M. Murabito, S. Bhasin, L. Vandenput, R. Haring,

Genetic determinants of serum testosterone concentrations in men. *PLoS Genet.* **7**, e1002313 (2011).

25. Z. Chen, S. Tao, Y. Gao, J. Zhang, Y. Hu, L. Mo, S.-T. Kim, X. Yang, A. Tan, H. Zhang, X. Qin, L. Li, Y. Wu, S. Zhang, S. L. Zheng, J. Xu, Z. Mo, J. Sun, Genome-wide association study of sex hormones, gonadotropins and sex hormone-binding protein in Chinese men. *J. Med. Genet.* **50**, 794–801 (2013).

26. L. Ellis, H. Nyborg, Racial/ethnic variations in male testosterone levels: A probable contributor to group differences in health. *Steroids*. **57** (1992), pp. 72–75.

27. D. S. Lopez, S. B. Peskoe, C. E. Joshu, A. Dobs, M. Feinleib, N. Kanarek, W. G. Nelson, E. Selvin, S. Rohrmann, E. A. Platz, Racial/ethnic differences in serum sex steroid hormone concentrations in US adolescent males. *Cancer Causes Control*. **24**, 817–826 (2013).

28. S. J. Nyante, B. I. Graubard, Y. Li, G. M. McQuillan, E. A. Platz, S. Rohrmann, G. Bradwin, K. A. McGlynn, Trends in sex hormone concentrations in US males: 1988-1991 to 1999-2004. *Int. J. Androl.* **35**, 456–466 (2012).

29. H. Fang, Q. Hui, J. Lynch, J. Honerlaw, T. L. Assimes, J. Huang, M. Vujkovic, S. M. Damrauer, S. Pyarajan, J. M. Gaziano, S. L. DuVall, C. J. O'Donnell, K. Cho, K.-M. Chang, P. W. F. Wilson, P. S. Tsao, VA Million Veteran Program, Y. V. Sun, H. Tang, Harmonizing Genetic Ancestry and Self-identified Race/Ethnicity in Genome-wide Association Studies. *Am. J. Hum. Genet.* **105**, 763–772 (2019).

30. B. K. Bulik-Sullivan, P.-R. Loh, H. K. Finucane, S. Ripke, J. Yang, Schizophrenia Working Group of the Psychiatric Genomics Consortium, N. Patterson, M. J. Daly, A. L. Price, B. M. Neale, LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).

31. F. Muntoni, S. Torelli, A. Ferlini, Dystrophin and mutations: one gene, several proteins, multiple phenotypes. *Lancet Neurol.* **2**, 731–740 (2003).

32. M. Lambert, P. Chafey, J. P. Hugnot, A. Koulakoff, Y. Berwald-Netter, C. Billard, G. E. Morris, A. Kahn, J. C. Kaplan, H. Gilgenkrantz, Expression of the transcripts initiated in the 62nd intron of the dystrophin gene. *Neuromuscul. Disord.* **3**, 519–524 (1993).

33. M. Sironi, A. Bardoni, G. Felisari, R. Cagliani, M. Robotti, G. P. Comi, M. Moggio, N. Bresolin, Transcriptional activation of the non-muscle, full-length dystrophin isoforms in Duchenne muscular dystrophy skeletal muscle. *J. Neurol. Sci.* **186**, 51–57 (2001).

34. R. D. Bies, D. Friedman, R. Roberts, M. B. Perryman, C. T. Caskey, Expression and localization of dystrophin in human cardiac Purkinje fibers. *Circulation*. **86**, 147–153 (1992).

35. B. Bulik-Sullivan, ReproGen Consortium, H. K. Finucane, V. Anttila, A. Gusev, F. R. Day, P.-R. Loh, L. Duncan, J. R. B. Perry, N. Patterson, E. B. Robinson, M. J. Daly, A. L. Price, B. M. Neale, Psychiatric Genomics Consortium, Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium, An atlas of genetic correlations across

human diseases and traits. *Nature Genetics*. **47** (2015), pp. 1236–1241.

36. J. Zheng, A. M. Erzurumluoglu, B. L. Elsworth, J. P. Kemp, L. Howe, P. C. Haycock, G. Hemani, K. Tansey, C. Laurin, Early Genetics and Lifecourse Epidemiology (EAGLE) Eczema Consortium, B. S. Pourcain, N. M. Warrington, H. K. Finucane, A. L. Price, B. K. Bulik-Sullivan, V. Anttila, L. Paternoster, T. R. Gaunt, D. M. Evans, B. M. Neale, LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics*. **33**, 272–279 (2017).

37. C. Giambartolomei, D. Vukcevic, E. E. Schadt, L. Franke, A. D. Hingorani, C. Wallace, V. Plagnol, Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).

38. M. D. M. Grasa, J. Gulfo, N. Camps, R. Alcalá, L. Monserrat, J. M. Moreno-Navarrete, F. J. Ortega, M. Esteve, X. Remesar, J. A. Fernández-López, J. M. Fernández-Real, M. Alemany, Modulation of SHBG binding to testosterone and estradiol by sex and morbid obesity. *Eur. J. Endocrinol.* **176**, 393–404 (2017).

39. H. Li, T. Pham, B. C. McWhinney, J. P. Ungerer, C. J. Pretorius, D. J. Richard, R. H. Mortimer, M. C. d'Emden, K. Richard, Sex Hormone Binding Globulin Modifies Testosterone Action and Metabolism in Prostate Cancer Cells. *Int. J. Endocrinol.* **2016**, 6437585 (2016).

40. A. D. Coviello, R. Haring, M. Wellons, D. Vaidya, T. Lehtimäki, S. Keildson, K. L. Lunetta, C. He, M. Fornage, V. Lagou, M. Mangino, N. C. Onland-Moret, B. Chen, J. Eriksson, M. Garcia, Y. M. Liu, A. Koster, K. Lohman, L.-P. Lyytikäinen, A.-K. Petersen, J. Prescott, L. Stolk, L. Vandenput, A. R. Wood, W. V. Zhuang, A. Ruokonen, A.-L. Hartikainen, A. Pouta, S. Bandinelli, R. Biffar, G. Brabant, D. G. Cox, Y. Chen, S. Cummings, L. Ferrucci, M. J. Gunter, S. E. Hankinson, H. Martikainen, A. Hofman, G. Homuth, T. Illig, J.-O. Jansson, A. D. Johnson, D. Karasik, M. Karlsson, J. Kettunen, D. P. Kiel, P. Kraft, J. Liu, Ö. Ljunggren, M. Lorentzon, M. Maggio, M. R. P. Markus, D. Mellström, I. Miljkovic, D. Mirel, S. Nelson, L. Morin Papunen, P. H. M. Peeters, I. Prokopenko, L. Raffel, M. Reincke, A. P. Reiner, K. Rexrode, F. Rivadeneira, S. M. Schwartz, D. Siscovick, N. Soranzo, D. Stöckl, S. Tworoger, A. G. Uitterlinden, C. H. van Gils, R. S. Vasan, H.-E. Wichmann, G. Zhai, S. Bhasin, M. Bidlingmaier, S. J. Chanock, I. De Vivo, T. B. Harris, D. J. Hunter, M. Kähönen, S. Liu, P. Ouyang, T. D. Spector, Y. T. van der Schouw, J. Viikari, H. Wallaschofski, M. I. McCarthy, T. M. Frayling, A. Murray, S. Franks, M.-R. Järvelin, F. H. de Jong, O. Raitakari, A. Teumer, C. Ohlsson, J. M. Murabito, J. R. B. Perry, A genome-wide association meta-analysis of circulating sex hormone-binding globulin reveals multiple Loci implicated in sex steroid hormone regulation. *PLoS Genet.* **8**, e1002805 (2012).

41. A. C. Pereira, J. D. Gray, J. F. Kogan, R. L. Davidson, T. G. Rubin, M. Okamoto, J. H. Morrison, B. S. McEwen, Age and Alzheimer's disease gene expression profiles reversed by the glutamate modulator riluzole. *Mol. Psychiatry*. **22**, 296–305 (2017).

42. S. Matsuda, Y. Matsuda, L. D'Adamio, BRI3 inhibits amyloid precursor protein processing

in a mechanistically distinct manner from its homologue dementia gene BRI2. *J. Biol. Chem.* **284**, 15815–15825 (2009).

43. N. Carrier, S. K. Saland, F. Duclot, H. He, R. Mercer, M. Kabbaj, The Anxiolytic and Antidepressant-like Effects of Testosterone and Estrogen in Gonadectomized Male Rats. *Biol. Psychiatry*. **78**, 259–269 (2015).

44. Y. Nakanishi, N. Akiyama, T. Tsukaguchi, T. Fujii, Y. Satoh, N. Ishii, M. Aoki, Mechanism of Oncogenic Signal Activation by the Novel Fusion Kinase FGFR3-BAIAP2L1. *Mol. Cancer Ther.* **14**, 704–712 (2015).

45. D.-Q. Guo, H. Zhang, S.-J. Tan, Y.-C. Gu, Nifedipine promotes the proliferation and migration of breast cancer cells. *PLoS One*. **9**, e113649 (2014).

46. I. Errazuriz, S. Dube, A. Basu, R. Basu, Effects of testosterone replacement on glucose and lipid metabolism. *Cardiovascular Endocrinology*. **4** (2015), pp. 95–99.

47. X. Huang, Z. Feng, Y. Jiang, J. Li, Q. Xiang, S. Guo, C. Yang, L. Fei, G. Guo, L. Zheng, Y. Wu, Y. Chen, VSIG4 mediates transcriptional inhibition of Nlrp3 and Il-1β in macrophages. *Science Advances*. **5** (2019), p. eaau7426.

48. K. Y. Helmy, K. J. Katschke Jr, N. N. Gorgani, N. M. Kljavin, J. M. Elliott, L. Diehl, S. J. Scales, N. Ghilardi, M. van Lookeren Campagne, CRIg: a macrophage complement receptor required for phagocytosis of circulating pathogens. *Cell*. **124**, 915–927 (2006).

49. L. Vogt, N. Schmitz, M. O. Kurrer, M. Bauer, H. I. Hinton, S. Behnke, D. Gatto, P. Sebbel, R. R. Beerli, I. Sonderegger, M. Kopf, P. Saudan, M. F. Bachmann, VSIG4, a B7 family-related protein, is a negative regulator of T cell activation. *J. Clin. Invest.* **116**, 2817–2826 (2006).

50. S. Tsukita, S. Yonemura, ERM (ezrin/radixin/moesin) family: from cytoskeleton to signal transduction. *Curr. Opin. Cell Biol.* **9**, 70–75 (1997).

51. A. Bretscher, K. Edwards, R. G. Fehon, ERM proteins and merlin: integrators at the cell cortex. *Nature Reviews Molecular Cell Biology*. **3** (2002), pp. 586–599.

52. P. S. Freymuth, H. L. Fitzsimons, The ERM protein Moesin is essential for neuronal morphogenesis and long-term memory in Drosophila. *Mol. Brain*. **10**, 41 (2017).

53. Y.-Q. Li, Z. Zheng, Q.-X. Liu, X. Lu, D. Zhou, J. Zhang, H. Zheng, J.-G. Dai, Moesin as a prognostic indicator of lung adenocarcinoma improves prognosis by enhancing immune lymphocyte infiltration. *World J. Surg. Oncol.* **19**, 109 (2021).

54. W. Liao, W. Huang, Y. Guo, M. Xin, X. Fu, Testosterone promotes vascular endothelial cell migration via upregulation of ROCK-2/moesin cascade. *Mol. Biol. Rep.* **40**, 6729–6735 (2013).

55. C. J. Malkin, P. J. Pugh, R. D. Jones, T. H. Jones, K. S. Channer, Testosterone as a protective

factor against atherosclerosis--immunomodulation and influence upon plaque development and stability. *J. Endocrinol.* **178**, 373–380 (2003).

56. L. Lanser, F. R. Burkert, L. Thommes, A. Egger, G. Hoermann, S. Kaser, G. M. Pinggera, M. Anliker, A. Griesmacher, G. Weiss, R. Bellmann-Weiler, Testosterone Deficiency Is a Risk Factor for Severe COVID-19. *Front. Endocrinol.* . **12**, 694083 (2021).

57. M. A. Rahim, M. Al Rahim, Y. Yoon, C. Dimovasili, Z. Shao, Q. Huang, E. Zhang, N. Kezunovic, L. Chen, A. Schaffner, G. W. Huntley, I. Ubarretxena-Belandia, A. Georgakopoulos, N. K. Robakis, Presenilin1 familial Alzheimer disease mutants inactivate EFNB1- and BDNF-dependent neuroprotection against excitotoxicity by affecting neuroprotective complexes of N-methyl-d-aspartate receptor. *Brain Communications*. **2** (2020), , doi:10.1093/braincomms/fcaa100.

58. A. Q. Nguyen, S. Sutley, J. Koeppen, K. Mina, S. Woodruff, S. Hanna, A. Vengala, P. W. Hickmott, A. Obenaus, I. M. Ethell, Astrocytic Ephrin-B1 Controls Excitatory-Inhibitory Balance in Developing Hippocampus. *The Journal of Neuroscience*. **40** (2020), pp. 6854–6871.

59. R. I. Aqeilan, J. P. Hagan, A. de Bruin, M. Rawahneh, Z. Salah, E. Gaudio, H. Siddiqui, S. Volinia, H. Alder, J. B. Lian, G. S. Stein, C. M. Croce, Targeted ablation of the WW domain-containing oxidoreductase tumor suppressor leads to impaired steroidogenesis. *Endocrinology*. **150**, 1530–1535 (2009).

60. B. W. Kunkle, B. Grenier-Boley, R. Sims, J. C. Bis, V. Damotte, A. C. Naj, A. Boland, M. Vronskaya, S. J. van der Lee, A. Amlie-Wolf, C. Bellenguez, A. Frizatti, V. Chouraki, E. R. Martin, K. Sleegers, N. Badarinarayan, J. Jakobsdottir, K. L. Hamilton-Nelson, S. Moreno-Grau, R. Olaso, R. Raybould, Y. Chen, A. B. Kuzma, M. Hiltunen, T. Morgan, S. Ahmad, B. N. Vardarajan, J. Epelbaum, P. Hoffmann, M. Boada, G. W. Beecham, J.-G. Garnier, D. Harold, A. L. Fitzpatrick, O. Valladares, M.-L. Moutet, A. Gerrish, A. V. Smith, L. Qu, D. Bacq, N. Denning, X. Jian, Y. Zhao, M. Del Zompo, N. C. Fox, S.-H. Choi, I. Mateo, J. T. Hughes, H. H. Adams, J. Malamon, F. Sanchez-Garcia, Y. Patel, J. A. Brody, B. A. Dombroski, M. C. D. Naranjo, M. Daniilidou, G. Eiriksdottir, S. Mukherjee, D. Wallon, J. Uphill, T. Aspelund, L. B. Cantwell, F. Garzia, D. Galimberti, E. Hofer, M. Butkiewicz, B. Fin, E. Scarpini, C. Sarnowski, W. S. Bush, S. Meslage, J. Kornhuber, C. C. White, Y. Song, R. C. Barber, S. Engelborghs, S. Sordon, D. Voijnovic, P. M. Adams, R. Vandenberghe, M. Mayhaus, L. A. Cupples, M. S. Albert, P. P. De Deyn, W. Gu, J. J. Himali, D. Beekly, A. Squassina, A. M. Hartmann, A. Orellana, D. Blacker, E. Rodriguez-Rodriguez, S. Lovestone, M. E. Garcia, R. S. Doody, C. Munoz-Fernadez, R. Sussams, H. Lin, T. J. Fairchild, Y. A. Benito, C. Holmes, H. Karamujić-Čomić, M. P. Frosch, H. Thonberg, W. Maier, G. Roshchupkin, B. Ghetti, V. Giedraitis, A. Kawalia, S. Li, R. M. Huebinger, L. Kilander, S. Moebus, I. Hernández, M. I. Kamboh, R. Brundin, J. Turton, Q. Yang, M. J. Katz, L. Concari, J. Lord, A. S. Beiser, C. D. Keene, S. Helisalmi, I. Kloszewska, W. A. Kukull, A. M. Koivisto, A. Lynch, L. Tarraga, E. B. Larson, A. Haapasalo, B. Lawlor, T. H. Mosley, R. B. Lipton, V. Solfrizzi, M. Gill, W. T. Longstreth Jr, T. J. Montine, V. Frisardi, M. Diez-Fairen, F. Rivadeneira, R. C. Petersen, V. Deramecourt, I. Alvarez, F. Salani, A. Ciaramella, E. Boerwinkle, E. M. Reiman, N. Fievet, J. I. Rotter, J. S. Reisch, O. Hanon, C. Cupidi, A. G.

Andre Uitterlinden, D. R. Royall, C. Dufouil, R. G. Maletta, I. de Rojas, M. Sano, A. Brice, R. Cecchetti, P. S. George-Hyslop, K. Ritchie, M. Tsolaki, D. W. Tsuang, B. Dubois, D. Craig, C.-K. Wu, H. Soininen, D. Avramidou, R. L. Albin, L. Fratiglioni, A. Germanou, L. G. Apostolova, L. Keller, M. Koutroumani, S. E. Arnold, F. Panza, O. Gkatzima, S. Asthana, D. Hannequin, P. Whitehead, C. S. Atwood, P. Caffarra, H. Hampel, I. Quintela, Á. Carracedo, L. Lannfelt, D. C. Rubinsztein, L. L. Barnes, F. Pasquier, L. Frölich, S. Barral, B. McGuinness, T. G. Beach, J. A. Johnston, J. T. Becker, P. Passmore, E. H. Bigio, J. M. Schott, T. D. Bird, J. D. Warren, B. F. Boeve, M. K. Lupton, J. D. Bowen, P. Proitsi, A. Boxer, J. F. Powell, J. R. Burke, J. S. K. Kauwe, J. M. Burns, M. Mancuso, J. D. Buxbaum, U. Bonuccelli, N. J. Cairns, A. McQuillin, C. Cao, G. Livingston, C. S. Carlson, N. J. Bass, C. M. Carlsson, J. Hardy, R. M. Carney, J. Bras, M. M. Carrasquillo, R. Guerreiro, M. Allen, H. C. Chui, E. Fisher, C. Masullo, E. A. Crocco, C. DeCarli, G. Bisceglio, M. Dick, L. Ma, R. Duara, N. R. Graff-Radford, D. A. Evans, A. Hodges, K. M. Faber, M. Scherer, K. B. Fallon, M. Riemenschneider, D. W. Fardo, R. Heun, M. R. Farlow, H. Kölsch, S. Ferris, M. Leber, T. M. Foroud, I. Heuser, D. R. Galasko, I. Giegling, M. Gearing, M. Hüll, D. H. Geschwind, J. R. Gilbert, J. Morris, R. C. Green, K. Mayo, J. H. Growdon, T. Feulner, R. L. Hamilton, L. E. Harrell, D. Drichel, L. S. Honig, T. D. Cushion, M. J. Huentelman, P. Hollingworth, C. M. Hulette, B. T. Hyman, R. Marshall, G. P. Jarvik, A. Meggy, E. Abner, G. E. Menzies, L.-W. Jin, G. Leonenko, L. M. Real, G. R. Jun, C. T. Baldwin, D. Grozeva, A. Karydas, G. Russo, J. A. Kaye, R. Kim, F. Jessen, N. W. Kowall, B. Vellas, J. H. Kramer, E. Vardy, F. M. LaFerla, K.-H. Jöckel, J. J. Lah, M. Dichgans, J. B. Leverenz, D. Mann, A. I. Levey, S. Pickering-Brown, A. P. Lieberman, N. Klopp, K. L. Lunetta, H.-E. Wichmann, C. G. Lyketsos, K. Morgan, D. C. Marson, K. Brown, F. Martiniuk, C. Medway, D. C. Mash, M. M. Nöthen, E. Masliah, N. M. Hooper, W. C. McCormick, A. Daniele, S. M. McCurry, A. Bayer, A. N. McDavid, J. Gallacher, A. C. McKee, H. van den Bussche, M. Mesulam, C. Brayne, B. L. Miller, S. Riedel-Heller, C. A. Miller, J. W. Miller, A. Al-Chalabi, J. C. Morris, C. E. Shaw, A. J. Myers, J. Wiltfang, S. O'Bryant, J. M. Olichney, V. Alvarez, J. E. Parisi, A. B. Singleton, H. L. Paulson, J. Collinge, W. R. Perry, S. Mead, E. Peskind, D. H. Cribbs, M. Rossor, A. Pierce, N. S. Ryan, W. W. Poon, B. Nacmias, H. Potter, S. Sorbi, J. F. Quinn, E. Sacchinelli, A. Raj, G. Spalletta, M. Raskind, C. Caltagirone, P. Bossù, M. D. Orfei, B. Reisberg, R. Clarke, C. Reitz, A. D. Smith, J. M. Ringman, D. Warden, E. D. Roberson, G. Wilcock, E. Rogaeva, A. C. Bruni, H. J. Rosen, M. Gallo, R. N. Rosenberg, Y. Ben-Shlomo, M. A. Sager, P. Mecocci, A. J. Saykin, P. Pastor, M. L. Cuccaro, J. M. Vance, J. A. Schneider, L. S. Schneider, S. Slifer, W. W. Seeley, A. G. Smith, J. A. Sonnen, S. Spina, R. A. Stern, R. H. Swerdlow, M. Tang, R. E. Tanzi, J. Q. Trojanowski, J. C. Troncoso, V. M. Van Deerlin, L. J. Van Eldik, H. V. Vinters, J. P. Vonsattel, S. Weintraub, K. A. Welsh-Bohmer, K. C. Wilhelmsen, J. Williamson, T. S. Wingo, R. L. Woltjer, C. B. Wright, C.-E. Yu, L. Yu, Y. Saba, A. Pilotto, M. J. Bullido, O. Peters, P. K. Crane, D. Bennett, P. Bosco, E. Coto, V. Boccardi, P. L. De Jager, A. Lleo, N. Warner, O. L. Lopez, M. Ingelsson, P. Deloukas, C. Cruchaga, C. Graff, R. Gwilliam, M. Fornage, A. M. Goate, P. Sanchez-Juan, P. G. Kehoe, N. Amin, N. Ertekin-Taner, C. Berr, S. Debette, S. Love, L. J. Launer, S. G. Younkin, J.-F. Dartigues, C. Corcoran, M. A. Ikram, D. W. Dickson, G. Nicolas, D. Campion, J. Tschanz, H. Schmidt, H. Hakonarson, J. Clarimon, R. Munger, R. Schmidt, L. A. Farrer, C. Van Broeckhoven, M. C O'Donovan, A. L. DeStefano, L. Jones, J. L. Haines, J.-F. Deleuze, M. J. Owen, V. Gudnason, R. Mayeux, V. Escott-Price, B. M. Psaty, A. Ramirez, L.-S. Wang, A. Ruiz, C. M. van Duijn, P. A. Holmans, S. Seshadri, J. Williams, P. Amouyel, G. D.

Schellenberg, J.-C. Lambert, M. A. Pericak-Vance, Alzheimer Disease Genetics Consortium (ADGC), European Alzheimer's Disease Initiative (EADI), Cohorts for Heart and Aging Research in Genomic Epidemiology Consortium (CHARGE),, Genetic and Environmental Risk in AD/Defining Genetic, Polygenic and Environmental Risk for Alzheimer's Disease Consortium (GERAD/PERADES), Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. *Nat. Genet.* **51**, 414–430 (2019).

61. L. Carcaillon, S. Brailly-Tabard, M.-L. Ancelin, C. Tzourio, A. Foubert-Samier, J.-F. Dartigues, A. Guiochon-Mantel, P.-Y. Scarabin, Low testosterone and the risk of dementia in elderly men: Impact of age and education. *Alzheimers. Dement.* **10**, S306–14 (2014).

62. E. J. Wahjoepramono, P. R. Asih, V. Aniwiyanti, K. Taddei, S. S. Dhaliwal, S. J. Fuller, J. Foster, M. Carruthers, G. Verdile, H. R. Sohrabi, R. N. Martins, The Effects of Testosterone Supplementation on Cognitive Functioning in Older Men. *CNS Neurol. Disord. Drug Targets*. **15**, 337–343 (2016).

63. C. E. Jaquish, J. Blangero, S. M. Haffner, M. P. Stern, J. W. MacCluer, Quantitative genetics of serum sex hormone—binding globulin levels in participants in the San Antonio Family Heart Study. *Metabolism*. **46** (1997), pp. 988–991.

64. H. Z. Ring, C. N. Lessov, T. Reed, R. Marcus, L. Holloway, G. E. Swan, D. Carmelli, Heritability of plasma sex hormones and hormone binding globulin in adult male twins. *J. Clin. Endocrinol. Metab.* **90**, 3653–3658 (2005).

65. V. Bogaert, Y. Taes, P. Konings, K. Van Steen, D. De Bacquer, S. Goemaere, H. Zmierczak, P. Crabbe, J.-M. Kaufman, Heritability of blood concentrations of sex-steroids in relation to body composition in young adult male siblings. *Clin. Endocrinol.* . **69**, 129–135 (2008).

66. S. Shifman, Linkage disequilibrium patterns of the human genome across populations. *Human Molecular Genetics*. **12** (2003), pp. 771–776.

67. J. J. Carrero, A. R. Qureshi, A. Nakashima, S. Arver, P. Parini, B. Lindholm, P. Bárány, O. Heimbürger, P. Stenvinkel, Prevalence and clinical implications of testosterone deficiency in men with end-stage renal disease. *Nephrol. Dial. Transplant*. **26**, 184–190 (2011).

68. G. Snyder, D. A. Shoskes, Hypogonadism and testosterone replacement therapy in end-stage renal disease (ESRD) and transplant patients. *Transl. Androl. Urol.* **5**, 885–889 (2016).

69. J. V. Zhao, C. M. Schooling, The role of testosterone in chronic kidney disease and kidney function in men and women: a bi-directional Mendelian randomization study in the UK Biobank. *BMC Med.* **18**, 122 (2020).

70. Y.-X. Chen, W. Zhang, W.-M. Wang, X.-L. Yu, Y.-M. Wang, M.-J. Zhang, N. Chen, Role of moesin in renal fibrosis. *PLoS One*. **9**, e112936 (2014).

71. C. J. Hodonsky, D. Jain, U. M. Schick, J. V. Morrison, L. Brown, C. P. McHugh, C. Schurmann, D. D. Chen, Y. M. Liu, P. L. Auer, C. A. Laurie, K. D. Taylor, B. L. Browning,

Y. Li, G. Papanicolaou, J. I. Rotter, R. Kurita, Y. Nakamura, S. R. Browning, R. J. F. Loos, K. E. North, C. C. Laurie, T. A. Thornton, N. Pankratz, D. E. Bauer, T. Sofer, A. P. Reiner, Genome-wide association study of red blood cell traits in Hispanics/Latinos: The Hispanic Community Health Study/Study of Latinos. *PLoS Genet.* **13**, e1006760 (2017).

72. M. R. Carnethon, J. Pu, G. Howard, M. A. Albert, C. A. M. Anderson, A. G. Bertoni, M. S. Mujahid, L. Palaniappan, H. A. Taylor Jr, M. Willis, C. W. Yancy, American Heart Association Council on Epidemiology and Prevention; Council on Cardiovascular Disease in the Young; Council on Cardiovascular and Stroke Nursing; Council on Clinical Cardiology; Council on Functional Genomics and Translational Biology; and Stroke Council, Cardiovascular Health in African Americans: A Scientific Statement From the American Heart Association. *Circulation*. **136**, e393–e423 (2017).

73. M. C. Marshall Jr, Diabetes in African Americans. *Postgrad. Med. J.* **81**, 734–740 (2005).

74. M. Laster, J. I. Shen, K. C. Norris, Kidney Disease Among African Americans: A Population Perspective. *Am. J. Kidney Dis.* **72**, S3–S7 (2018).

75. S. B. Nicholas, K. Kalantar-Zadeh, K. C. Norris, Racial Disparities in Kidney Disease Outcomes. *Seminars in Nephrology*. **33** (2013), pp. 409–415.

76. K. Norris, A. R. Nissenson, Race, gender, and socioeconomic disparities in CKD in the United States. *J. Am. Soc. Nephrol.* **19**, 1261–1270 (2008).

77. M. I. Harris, K. M. Flegal, C. C. Cowie, M. S. Eberhardt, D. E. Goldstein, R. R. Little, H.-M. Wiedmeyer, D. D. Byrd-Holt, Prevalence of Diabetes, Impaired Fasting Glucose, and Impaired Glucose Tolerance in U.S. Adults: The Third National Health and Nutrition Examination Survey, 1988-1994. *Diabetes Care*. **21** (1998), pp. 518–524.

78. R. Bell, *Diabetes 2001 Vital Statistics* (American Diabetes Association, 2003).

79. U. S. D. of H. &. H. S. A. F. H. R. A. Quality, US Department of Health & Human Services; Agency for Healthcare Research and Quality, National Healthcare Disparities Report. *PsycEXTRA Dataset* (2003), , doi:10.1037/e308932005-001.

80. L. Dong, O. A. Fakeye, G. Graham, D. J. Gaskin, Racial/Ethnic Disparities in Quality of Care for Cardiovascular Disease in Ambulatory Settings: A Review. *Med. Care Res. Rev.* **75**, 263–291 (2018).

81. G. H. Gibbons, Physiology, Genetics, and Cardiovascular Disease: Focus on African Americans. *The Journal of Clinical Hypertension*. **6** (2004), pp. 11–18.

82. S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, P. C. Sham, PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*. **81** (2007), pp. 559–575.

83. J. Knight, S. L. Spain, F. Capon, A. Hayday, F. O. Nestle, A. Clop, Wellcome Trust Case

Control Consortium, Genetic Analysis of Psoriasis Consortium, I-chip for Psoriasis Consortium, J. N. Barker, M. E. Weale, R. C. Trembath, Conditional analysis identifies three novel major histocompatibility complex loci associated with psoriasis. *Hum. Mol. Genet.* **21**, 5185–5192 (2012).

84. H. K. Finucane, ReproGen Consortium, B. Bulik-Sullivan, A. Gusev, G. Trynka, Y. Reshef, P.-R. Loh, V. Anttila, H. Xu, C. Zang, K. Farh, S. Ripke, F. R. Day, S. Purcell, E. Stahl, S. Lindstrom, J. R. B. Perry, Y. Okada, S. Raychaudhuri, M. J. Daly, N. Patterson, B. M. Neale, A. L. Price, Schizophrenia Working Group of the Psychiatric Genomics Consortium, The RACI Consortium, Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics*. **47** (2015), pp. 1228–1235.

85. A. Gusev, A. Ko, H. Shi, G. Bhatia, W. Chung, B. W. J. H. Penninx, R. Jansen, E. J. C. de Geus, D. I. Boomsma, F. A. Wright, P. F. Sullivan, E. Nikkola, M. Alvarez, M. Civelek, A. J. Lusis, T. Lehtimäki, E. Raitoharju, M. Kähönen, I. Seppälä, O. T. Raitakari, J. Kuusisto, M. Laakso, A. L. Price, P. Pajukanta, B. Pasaniuc, Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252 (2016).

86. Y. Benjamini, Y. Hochberg, Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*. **57** (1995), pp. 289–300.

87. S. Seabold, J. Perktold, Statsmodels: Econometric and Statistical Modeling with Python. *Proceedings of the 9th Python in Science Conference* (2010), , doi:10.25080/majora-92bf1922-011.

3.2.1 Abstract

Prostate cancer (PrCa) is one of the most genetically driven solid cancers with heritability estimates as high as 57%. African American men are at an increased risk of PrCa; however, current risk prediction models are based on European ancestry groups and may not be broadly applicable. In this study, we define an African ancestry group of 4,533 individuals to develop an African ancestry-specific PrCa polygenic risk score (PRState). We identified risk loci on chromosomes 3, 8, and 11 in the African ancestry group GWAS and constructed a polygenic risk score (PRS) from 10 African ancestry-specific PrCa risk SNPs, achieving an AUC of 0.61 [0.60-0.63] and 0.65 [0.64-0.67], when combined with age and family history. Performance dropped significantly when using ancestry-mismatched PRS models but remained comparable when using trans-ancestry models. Importantly, we validated the PRState score in the Million Veteran Program, demonstrating improved prediction of PrCa and metastatic PrCa in African American individuals. This study underscores the need for inclusion of individuals of African ancestry in gene variant discovery to optimize PRS.

3.2.2 Introduction

Prostate cancer (PrCa) remains the most common non-skin malignancy in men, with significant mortality resulting in 1 in 42 men diagnosed with PrCa dying from the disease(*1*, *2*). Men with an African ancestry have a 1.6 and 2.4 fold increased risk of PrCa diagnosis and age-matched mortality compared to men with a European ancestry(*3*, *4*). Multiple studies suggest genetic heritability is high for PrCa(*5*, *6*), with twin studies attributing 57% of PrCa risk to genetic factors(*7*).

While rare high penetrance genes and missense mutations (e.g., G84E in HOXB13) have been described, they represent an exceedingly small minority of PrCa cases. Single nucleotide polymorphisms (SNPs) in non-coding regions also contribute to PrCa risk, with many falling in the chromosome 8q24 risk region(*8*, *9*). Genome-wide association studies (GWAS) have identified more than 260 of these SNP susceptibility loci(*10*, *11*). However, the majority of discovery populations in these GWAS have been of European or Asian ancestry and studies on the role of ancestral genetic background in PrCa risk for other ethnic groups are needed(*8*, *10*, *12*, *13*).

Incorporating PrCa risk SNPs into a meaningful clinical tool is possible with polygenic risk scores (PRSs) that predict PrCa risk based on the presence of individual inherited SNPs(*14–16*). Sun et al demonstrated that addition of a PRS to family history improved the performance of predicting PrCa in populations of predominantly European ancestry(*11*, *16–20*). The full utility of such tools for diverse populations or in combination with nomograms is yet to be realized.

Here we use genetic ancestry to separate the ELLIPSE consortium into ancestry groups. We ran association studies within our African ancestry group to identify population-specific SNPs. We then constructed an African ancestry-specific PrCa PRS (PRState) that achieved an AUC of 0.65 [0.64-0.67] when combined with family history and age of diagnosis. Efficacy of PRS was contingent on inclusion of African ancestry group individuals in PRS construction. When only European ancestry group individuals were used to construct PRS, a considerable drop in performance was observed within the African ancestry group. PRS construction from trans-ancestry groups performed comparably to PRState.

Variants in PRState score have been described in previous trans-ancestry analysis; however, we demonstrate these variants contribute to a boost in PrCa prediction performance in both the ELLIPSE and Million Veteran Program African ancestry groups. These findings highlight

the importance of ancestry-specific risk SNP identification and will hopefully guide future PRS studies of PrCa in African ancestry groups.

3.2.3 Results

3.2.3.1 European and African Ancestry Group Identification in ELLIPSE

PCA analysis of 55 ancestry informative markers (AIMs) proposed by Kidd et al(*25*) revealed that individuals in the ELLIPSE could be stratified according to ancestral background. European, African and Asian descent individuals formed distinct clusters **(Figure 3.8)**. European and African ancestry likelihood thresholds were selected such that population size was maximized while minimizing admixture. The European ancestry group included individuals with European ancestry likelihood ratio > -10, resulting in 5,567 individuals. The African ancestry group included individuals with African ancestry likelihood ratio > -15 and European ancestry Likelihood ratio < -15, resulting in 4,533 individuals **(Figure 3.8)**. PCA analysis confirmed thresholds results in tightly clustered European and African ancestry groups. Self-identified ancestry aligned with genetic ancestry defined through AIMs; although some individuals self-identifying as Hispanic were included in European and African ancestry groups (**Table 3.3)**.

3.2.3.2 Inclusion of African Ancestry Group Individuals in Polygenic Risk Score Construction Improves African Ancestry Group Prostate Cancer Prediction

GWAS was performed in the African ancestry group to identify African ancestry-specific risk loci. Loci on chromosomes 3, 8, and 11 were significantly associated with PrCa risk in African American men **(Figure 3.9)**. Age, family history, genetic risk score and the combination of these 3 factors for PrCa were evaluated using ROC-AUC analysis. Risk prediction models using genetics

resulted in an AUC of 0.61 [0.60-0.63] **(Figure S3.6)** while a combined model (genetics, age and family history) resulted in an AUC of 0.65 **(Figure 3.9)**. To determine the efficacy of using a matched ancestry group model for prediction, we evaluated performance of a European ancestry group model on prediction of PrCa in our African ancestry group. Interestingly, model performance was poor with near random performance using genetics [AUC: 0.52 (0.50-0.53)] (**Figure S3.6)**. When genetics were combined with age and family history, the AUC was 0.59 [0.57-0.60], significantly lower than results from using the matched ancestry group model (**Figure 3.9)**.

Lastly, recent studies of trans-ancestral analysis of PrCa risk have been conducted(*11*), so we wanted to evaluate the performance of using a trans-ancestry group model on our African ancestry group. We trained models on a training set composed of both European and African ancestry groups combined and tested on the African ancestry group. We achieved comparable performance to matched ancestry group models. We achieved an AUC of 0.62 [0.61-0.64] using only genetics **(Figure S3.6)** and 0.66 [0.65-0.68] when combined with age and family history **(Figure 3.9)**. Individuals in the top 10th quantile of PRS constructed from matched ancestry model had 2-fold greater risk of PrCa compared to the 50th quantile (**Figure 3.9).** Quantile analysis with odds of PrCa using a European ancestry group model demonstrates no trend between PRS and PrCa risk in the African American ELLIPSE Consortium (**Figure 3.9).**

Since the trans-ancestry group model used 5 variants which overlapped with the African ancestry group model, we compared odds ratios of African ancestry-specific risk variants in European, African and trans-ancestry groups to determine if variants exhibited different associations with PrCa. Odds ratios between trans-ancestry group and African ancestry group association analyses were similar, compared to European ancestry group association analysis

(Figure S3.7). Odds ratios for African ancestry group individuals for 3 of the 8 chromosome 8 variants included in PRS construction were significantly different compared to European ancestry group individuals. These results demonstrate that PrCa variants have different effects based on ancestral background.

3.2.3.3 African Ancestry PrCa Risk Variants Improve Prostate Cancer Prediction when Combined with Previous PrCa Risk Variants

After demonstrating the importance of inclusion of ancestry-matched individuals in polygenic risk score prediction, we sought to evaluate how our PRS constructed with 10 African ancestry-specific variants, which we will now refer to as PRState, performed in comparison to previous models. We compared performance to a PRS from a previously published large trans-ancestry analysis of PrCa by Conti et al.(*11*). Five of the 10 African ancestry-specific variants we identified were in high linkage disequilibrium with variants previously implicated by Conti et al. and 2 of these 5 variants passed genome-wide significance threshold in Conti et al. African ancestry group GWAS. We compared our PRState score with the Conti et al. PRS constructed excluding these 5 variants. Addition of PRState score to Conti et al. PRS significantly improves prediction of PrCa by itself (DeLong P < .0003) (**Figure 3.10)** and with family history and age (DeLong P < .0003) **(Figure 3.10)**.

To validate our results, we compared PRState score and Conti et al. PRS performance in the Million Veteran Program. Average AUC for PrCa **(Figure 5A)** (DeLong P < 1e-16) and metastatic PrCa **(Figure 3.11)** (DeLong P < 8e-06) prediction was significantly higher in individuals of African ancestry when PRState and Conti et al PRS were combined compared to Conti et al. PRS alone. Combined PRState and Conti score was not associated with significantly

higher AUC in predicting death from PrCa **(Figure 3.11)** (DeLong P < .67) . Interestingly, we find that PRState score performance was associated with significantly better predictive value compared to European individuals for all three defined clinical PrCa endpoints **(Figure 3.12)** (PrCa DeLong P < 1e-16, metastatic PrCa DeLong P < 1e-16, fatal PrCa DeLong P < 1e-16). Furthermore, we characterized odds ratios of PRState variants between African and European HARE groups in the Million Veteran Program and noted a significant difference in the odds ratios for certain variants in all three endpoints (PrCa diagnosis, metastasis and death) (**Figure S3.8).** These results not only demonstrate the need to include individuals of African ancestry in construction of polygenic risk scores that predict PrCa risk, but also that African ancestry-specific variants are critical for prediction of other PrCa characteristics, such as metastasis.

3.2.3 Discussion

A critical limitation in most genetic studies in PrCa has been the overrepresentation of men with non-Hispanic European ancestry. Considering both the higher incidence and mortality of PrCa in African American men, this problem prevents the discovery of gene variants conferring PrCa risk in African and other ancestries. Using PrCa risk SNPs identified to be specific for African American men in ELLIPSE consortium, we used variants on chromosomes 3, 8, and 11 to construct an African ancestry-specific polygenic score (PRState) that achieved AUC of 0.61 [0.60-0.63] and 0.65 [0.64-0.67], when family history and age were added. We then compared PRState performance to models constructed from a primarily European ancestry group and a mixed European and African ancestry (trans-ancestry) group. We achieved comparable performance using a trans-ancestry group model, but a drop in performance using a European ancestry group model in ELLIPSE. The PRState score improved PrCa prediction performance when combined

with a PRS constructed from a previous larger PrCa meta-analysis conducted by Conti et al.(*11*) Although half of the PRState variants were in high linkage disequilibrium with Conti et al. variants, we demonstrate these African ancestry-specific variants significantly improve PrCa prediction in our discovery cohort (ELLIPSE) and external validation cohort (Million Veteran Program) of African ancestry group individuals. The results of our PRState study underscore the importance of including men with African ancestry when building genetic risk models and highlight African ancestry-specific PrCa variants that warrant further investigation.

PrCa is one of the most heritable cancer types and ancestry is an important determinant of PrCa risk. Using only a 55 SNP Panel and likelihood estimates from forensic genetic tool FROG-Kb, we were able to estimate genetic ancestry that aligned with self-identified ancestry in the ELLIPSE consortium. Although we had self-identified ancestry information available, our approach could be tested in cohorts where self-identified ancestry was not acquired. FROG-kb returns ancestry likelihood estimates for any panel of populations and individuals could have high ancestry likelihood estimates for several groups. For our study, we used FROB-kb to define categorical ancestry groups for PrCa risk variant discovery in the African ancestry group. Specifically, for defining the African ancestry group, we used a low European likelihood threshold combined with a high African likelihood threshold to define groups with little overlap in principal component analysis (**Figure 3.8).** However, FROB-kb estimates do not have to be used categorically and can also be incorporated in the model. Specifically, for admixed individuals, this quantitative estimate of ancestry likelihood estimates may be useful in genetic risk models.

Despite a relatively small discovery cohort of 4,533 individuals, we were able to identify 10 African ancestry-specific variants (of which 5 were novel), that predicted PrCa in ELLIPSE similar to the PRS constructed from Conti et al. from over 250 variants (ELLIPSE PRState AUC

= 0.66 [0.64-0.67] versus Conti PRS AUC = 0.69 [0.67-0.70]). Half of the PRState variants overlapped with variants identified by Conti et al., which included over 5x more individuals of African ancestry than our discovery cohort. Half of the PRState variants had been reported by Conti et al, whereas the novel variants included in PRState were identified using our methods. We applied a rigorous selection process for an African ancestry group that allowed identification of PrCa variants with African ancestry-specific effects. In order to test if our small sample was overfitting a prediction model, we applied our findings to an external cohort (the Million Veteran Program). PRState improved prediction of PrCa detection and demonstrated these variants were critical for prediction of high-risk PrCa that leads to metastasis in an African ancestry group within the Million Veteran Program. PRState was also associated with significantly higher AUC in the African ancestry group compared to European ancestry group for PrCa, metastatic PrCa and fatal PrCa. These results suggest PRState improves prediction of metastatic prostate cancer, which may help to prevent overtreatment of clinically non-aggressive prostate cancer with good survival

In the above analysis, we found PrCa risk SNPs in men ofAfrican descent are located at distinct loci that differ from PrCa risk loci identified in men of European ancestry. Variants on chromosome 8 were identified in both European and African ancestry group PrCa GWAS, however, the African ancestry-specific chromosome 8 variants (rs113343238, rs16902008, rs943270004, rs116845582, rs59825493) were not significantly associated with PrCa risk in the European ancestry group in the ELLIPSE Consortium(*26*, *27*). Interestingly, in the Million Veteran Program, certain variants were protective in the European ancestry group but associated with higher PrCa diagnosis in the African ancestry group (rs113343238, rs943270004). These results suggest that even within well-known PrCa risk loci, defining ancestry group differences will likely improve genetic risk models.

Our PRState analyses identified 10 African ancestry-specific variants in our discovery cohort of 4,533 individuals and demonstrated its improved predictive power. With a larger cohort, we could apply our approach to identify potential novel African ancestry-specific PrCa risk variants. While this work has demonstrated the feasibility of using a small number of SNPs to define ancestry backgrounds and then predict genetic risk of PrCa, there are several limitations. The ELLIPSE data set is the largest complete PrCa cohort and we focused on 4,533 patients in an African ancestry group representing <5% of the total ELLIPSE cohort. Thus the study was not powered to identify potentially meaningful SNPs, as indicated by the suggestive peaks in the African ancestry group GWAS on chromosomes 9, and 12, which did not reach statistical significance. Additionally, we recognize African ancestry encompasses a wide breadth of genetic diversity that can not be wholly defined by a small sample. Nevertheless, we believe our PRState study indicates the inclusion of ancestral inherited risk is an important variable for analyzing PrCa risk similar to family history. Further investigation of a larger African ancestry sample will likely improve the signal and determine the magnitude of PrCa risk conferred by the African ancestry-specific SNPs identified in our study. Additionally, while one unintentional risk of PrCa treatment is overtreatment of low grade PrCa disease, the ELLIPSE data does not include high grade versus low grade cancer characterization. Therefore, we are unable to identify whether these SNPs confer higher risk for PrCa metastasis and death, although our analysis within the MVP demonstrates an association. Future studies will expand these methods to other under-represented ancestry groups with a goal of developing PrCa risk stratifying tools based on an individual's ancestral background. Increasing the number of non-white patients in databases such as the ELLIPSE consortium is a key element to furthering research in these groups.

3.2.4 Materials and Methods

3.2.4.1 ELLIPSE Study Subjects and Genotype

The Elucidating Loci Involved in Prostate Cancer Susceptibility (ELLIPSE) consortium prostate cancer meta-analysis and genotypes (dbGaP Study Accession: phs001120.v1.p1) was accessed to analyze Affymetrix genotype calls for 91,644 male PrCa case/controls.

3.2.4.2 Quality Assurance

PLINK (RRID:SCR_001757) genotype files consisting of 505,219 calls from the following consent groups were compiled: c1-c3,c6,c8,c10-18,c20,c23,c25,c27-28. Pre-imputation processing of autosomal and X chromosome genotypes followed below steps:

1. Duplicated variants were removed.
2. Heterozygous haploid SNPs were set to missing.
3. SNPs with call rate <90% were removed
4. SNPs with minor allele frequency (MAF) <1% were removed
5. Individuals with genotype coverage <90% were removed
6. Non-ACGT variants were removed.

Strand flips were reversed using snpflip. After preprocessing genotypes, the remaining 410,116 SNPs and 91,644 individuals were input to the secure Michigan Imputation Server (RRID:SCR_017579)(*21*). Whole-genome SNPs were imputed with Minimac4 (RRID:SCR_009292) and ancestry-matched reference panel 1000 Genomes Project Phase 3 version 5 (RRID:SCR_008801). Finally, post-imputation duplicated SNPs and SNPs with MAF <1% were removed.

3.2.4.3 Ancestry Likelihood Calculation (FROG-kb)

For ancestry group calculations, we opted for an ancestry group prediction tool that does not require relationships with other individuals, like PCA. FROG-kb(*22*) uses Kidd AISNP panel (55 SNPs) to predict likelihood ratios for world geographic regions. Likelihood ratios for 160 populations were calculated and averaged. European American likelihood ratios were determined from populations in "Europe" region and African American likelihood ratios were determined from populations in "African" region. For the final European ancestry group, we used a European log likelihood > -10, resulting in 5567 individuals. For the final African ancestry group, we used a European log likelihood < -15 and African log likelihood > -15, resulting in 4533 individuals.

3.2.4.4 Genome-wide association analyses (GWAS)

PLINK (RRID:SCR_001757) GLM method(*23*) was used to conduct association analyses with PrCa case/control in European and African ancestry groups. All associations were adjusted for the first 10 principal components (PCA with 55-SNP Kidd panel) and age.

3.2.4.5 Polygenic Risk Score Calculation

Association analyses within European, African or mixed ancestry training sets were conducted. Significant variants were identified through PLINK (RRID:SCR_001757) linkage-based clumping using a p1 threshold of 5e-08, a p2 threshold of 1e-05, an r2 threshold of 0.1 and a kb threshold of 1000 kb. Ten, seven and fourteen significant variants were identified through African, European and trans-ancestry analysis, respectively. For PRS construction, variants were weighted by log (base 10) odds ratio from the training set association statistics, oriented to PrCa

risk allele, and combined. ROC-AUC evaluation across folds was conducted using polygenic scores as predictions.

For mismatched ancestry group analysis, European ancestry training sets were used for prediction on African ancestry test sets. For trans-ancestry group analysis, European and African ancestry training sets were combined and tested on African ancestry test sets. All three ancestry group PRSs were evaluated using 10-fold cross validation. AUC for each fold and overall are reported. Confidence intervals were calculated using pROC R package.

For contextualization of our results in relation to other genetic risk models, we compared our PRState score (composed of 10 African ancestry-specific variants) to variants published recently in a large meta-analysis of prostate cancer by Conti et al(*11*). Half (5) of the variants used in the PRState score were in high linkage disequilibrium (r2 > 0.3) with Conti et al variants(*11*). We excluded these variants and constructed a Conti PRS with reported odds ratio for the African group. To evaluate PRState and Conti polygenic risk scores, we used both scores as features for a logistic regression model with default parameters. Predicted probabilities were used in ROC evaluation. For the Million Veteran Program (MVP), genotype dosages were extracted for 10 PRState variants across participants and weighted by log (base 10) odds ratio from the best-performing fold in ELLIPSE. Performance was evaluated using ROC-AUC analysis in European and African ancestry groups.

Individual ancestry groups in the Million Veteran Program were characterized through Harmonized Ancestry and Race/Ethnicity (HARE) grouping(*24*). HARE grouping was specifically developed to categorize MVP individuals based on self-reported ancestry and genetic ancestry. HARE utilizes a support vector machine to output probabilities of an individual's ancestry group using self-identified and genetic ancestry. PrCa, metastatic PrCa and fatal PrCa

status was determined through ICD 9/10 diagnosis, procedure code, CPT and HCPCS procedure code, laboratory values, medications and clinical notes from inpatient, outpatient and fee-based care in the VA healthcare system. Family history information was available for only 55,610 of 121,964 African individuals and 322,706 of 461,627 European individuals in MVP. When evaluating genetic information only, the full population was used and the subset of the population with family history information available was used for additional multivariable association. ROC-AUC analysis was conducted in European and African ancestry individuals separately.

3.2.5 Figures

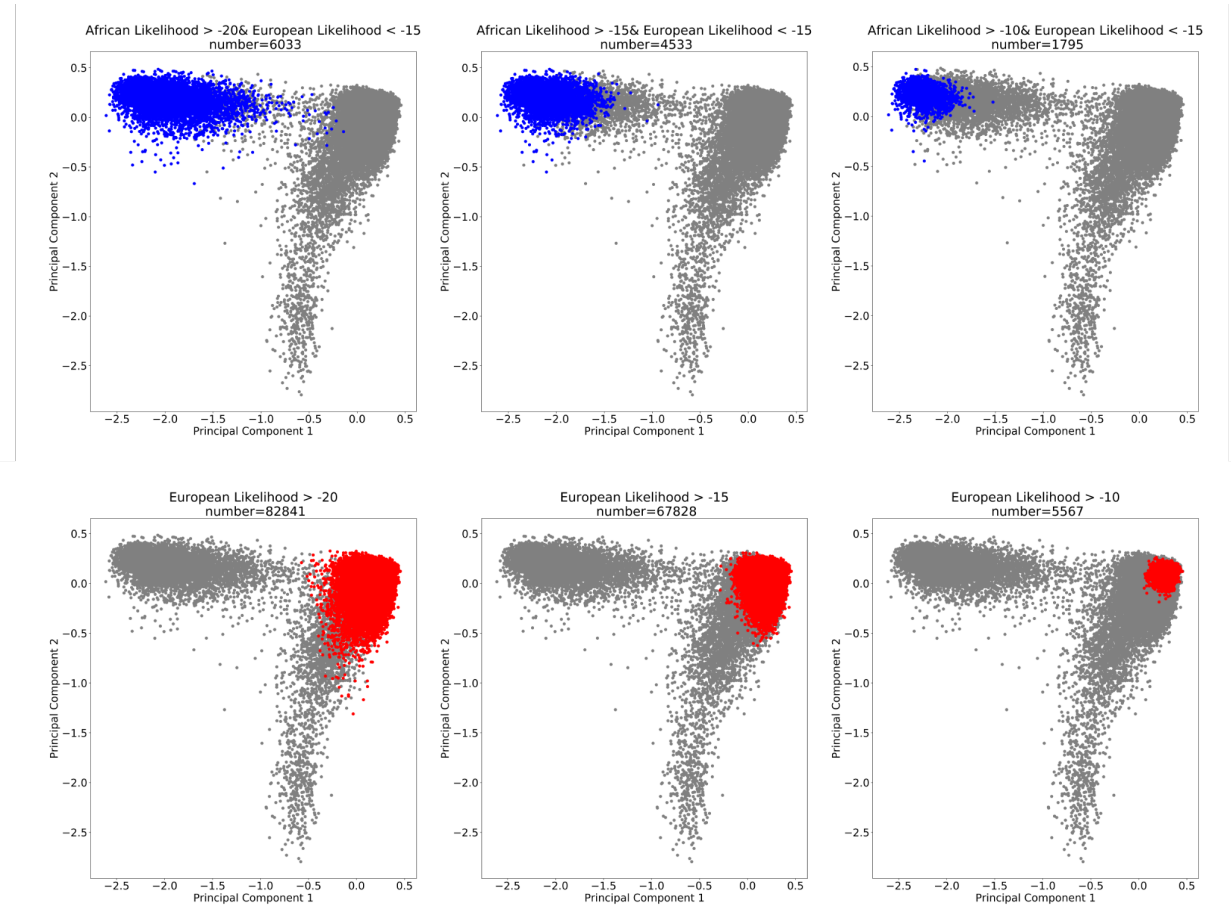

Figure 3.8 FROG-kb Ancestral Group selection PCA analysis was conducted with Kidd et al. ancestry informative markers (AIMs). Identified groups were determined from European and African ancestry likelihood ratios (FROG-kb, Kidd et al. panel).

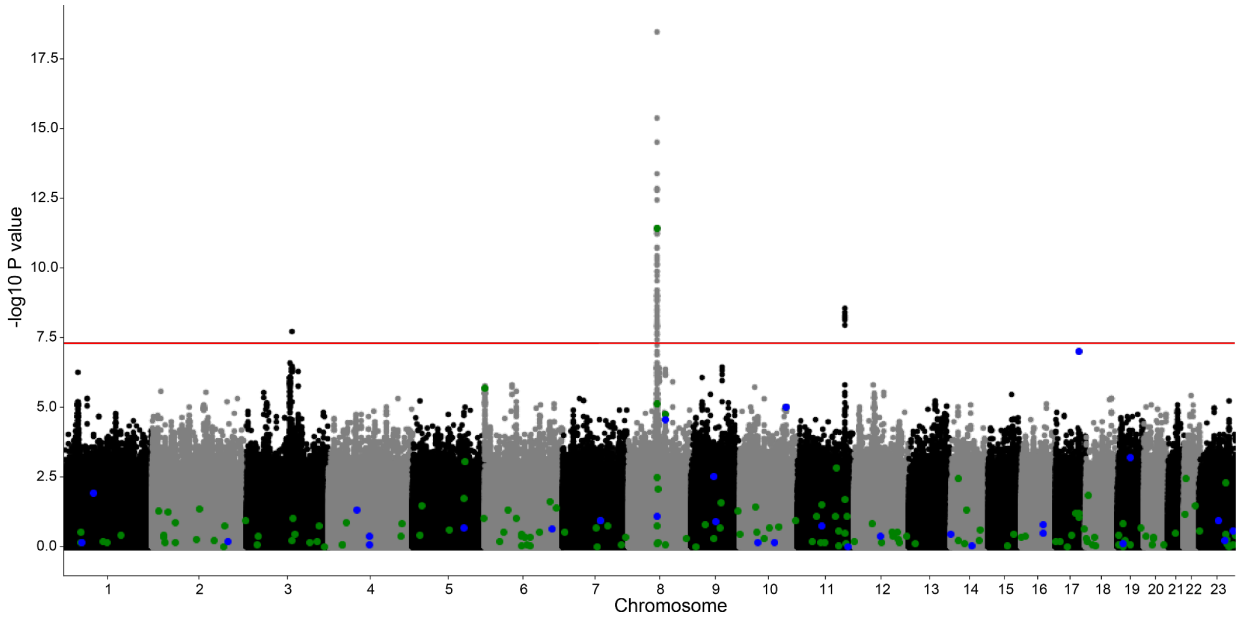

Figure 3.9 Manhattan Plot of Prostate Cancer (PrCa) Risk Manhattan plot of logistic association of genetic variants with PrCa risk in ELLIPSE African ancestry group (n=4,533). green=known PrCa risk SNPs, blue=SNPs associated with PSA levels

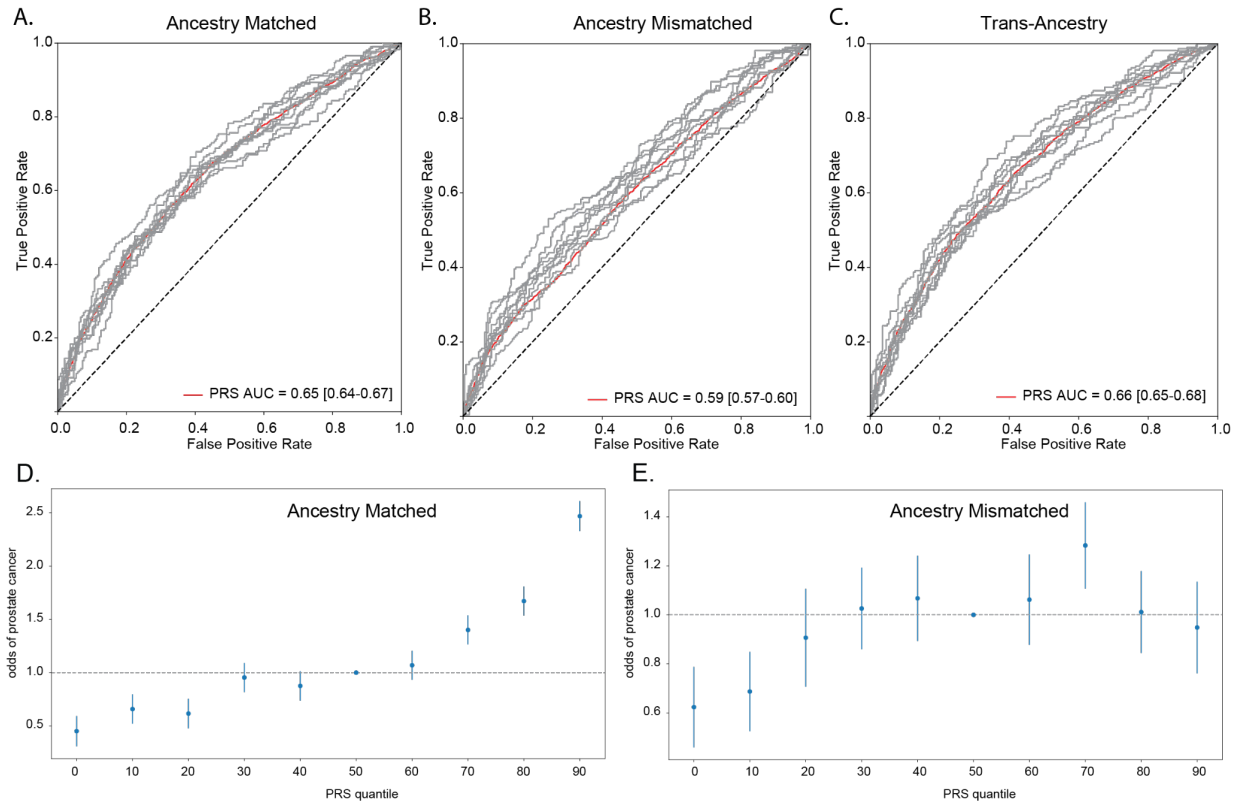

Figure 3.10 Performance of Polygenic Risk Scores (PRSs) Constructed from Different Ancestral Backgrounds in African Ancestry Group in ELLIPSE Consortium ROC curve for genetic prediction of PrCa risk in ELLIPSE Consortium African ancestry group (n=4,533) using: **(A)** PRSs constructed from 10 African ancestry-specific variants with age and family history. **(B)** PRSs constructed from 7 European ancestry-specific variants with age and family history. **(C)** PRSs constructed from 14 trans-ancestry specific variants with age and family history. Quantile plot of PRS constructed from: 10 African ancestry-specific variants **(D)** and 7 European ancestry-specific variants **(E)** and respective odds of prostate cancer.

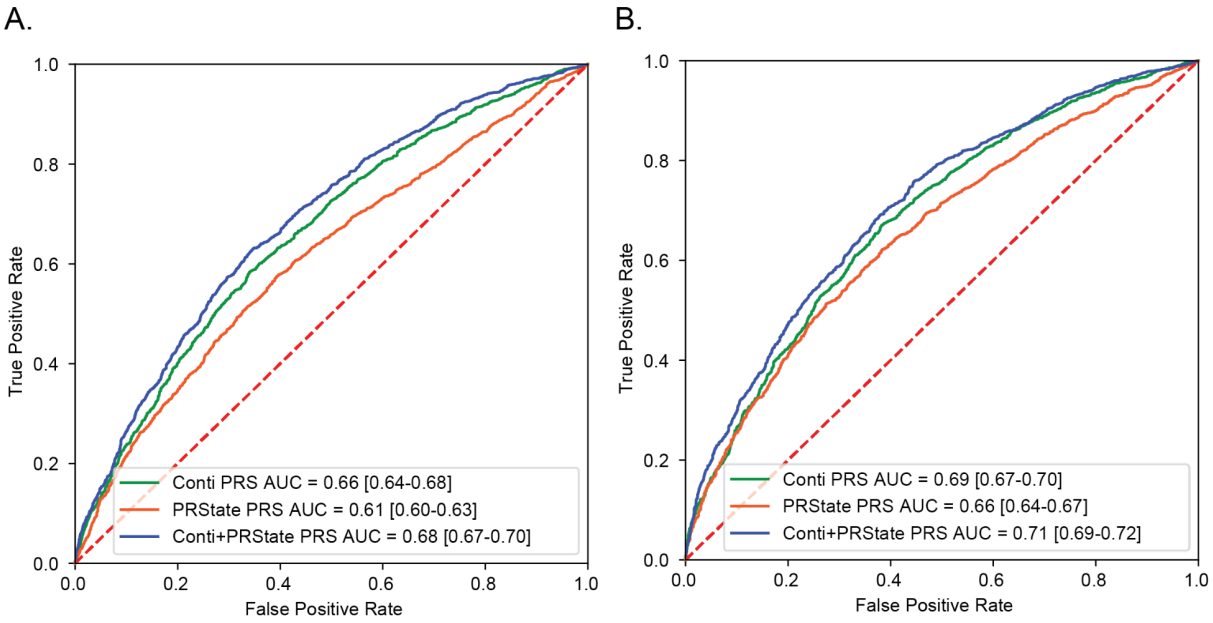

Figure 3.11 Evaluation of PRState and Conti Score in ELLIPSE Consortium (A) ROC curve for genetic prediction of PrCa risk in ELLIPSE Consortium African ancestry group (n=4,533) using Conti PRS, PRState PRS and combined Conti and PRState PRS (DeLong $p < 0.0003$). **(B)** ROC curve for genetic prediction of PrCa risk in ELLIPSE Consortium African ancestry group (n=4,533) using Conti PRS, PRState PRS and combined Conti and PRState PRS along with family history and age (DeLong $p < 0.0003$).

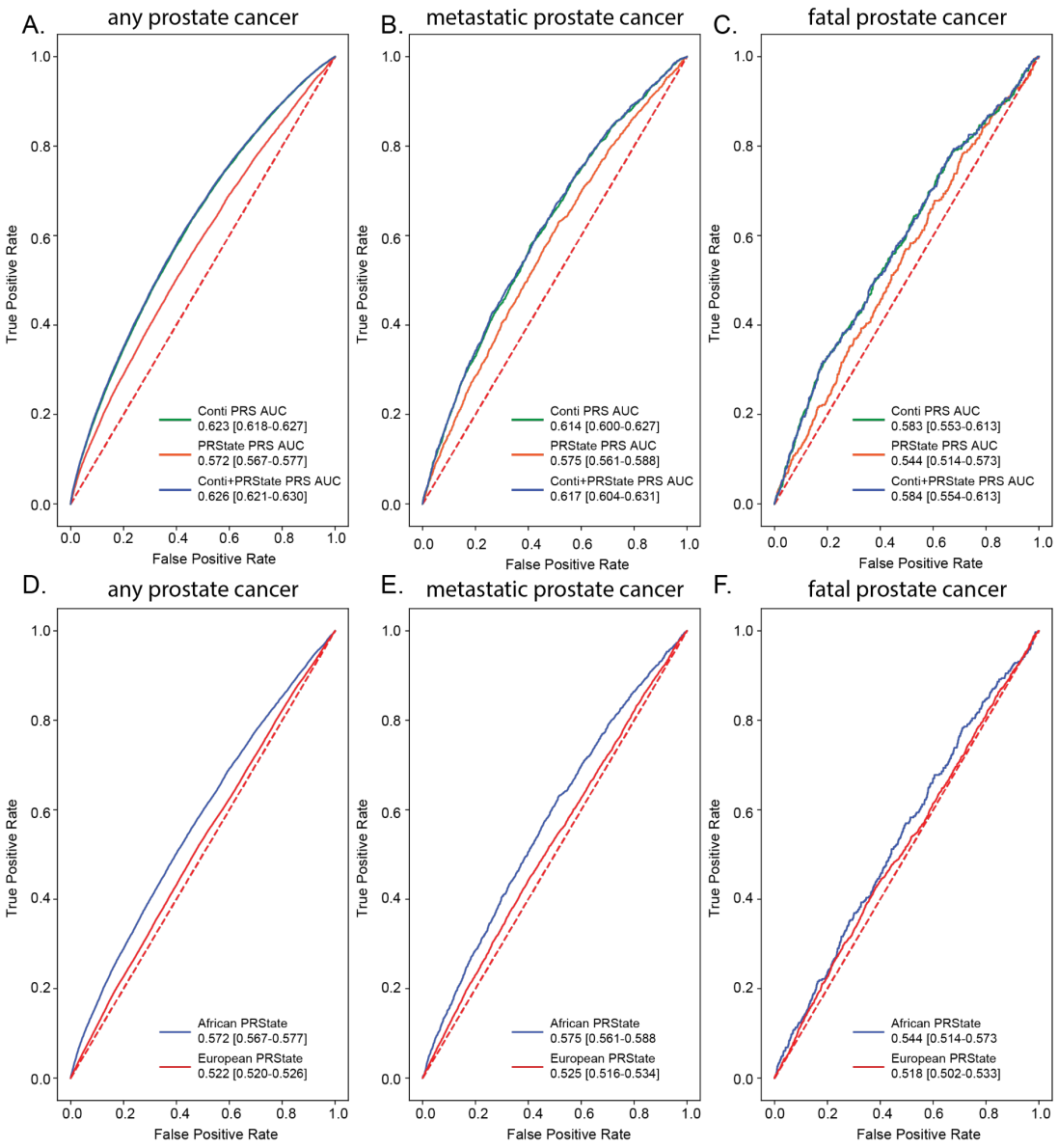

Figure 3.12 Evaluation of PRState and Conti Score in Million Veteran Program ROC curve for genetic prediction of prostate cancer risk in Million Veteran Program African HARE group (n=121,964) using Conti PRS, PRState PRS and combined Conti and PRState PRS for any prostate cancer (**A),** metastatic prostate cancer (**B),** and fatal prostate cancer (**C**). ROC curve for genetic prediction of prostate cancer risk in Million Veteran Program African HARE group (n=121,964) and European HARE group (n=461,627) using Conti PRS, PRState PRS and combined Conti and PRState PRS for any prostate cancer (**D),** metastatic prostate cancer (**E),** and fatal prostate cancer (**F**).

3.2.6 Tables

Table 3.3 Self-Identified Ancestry of Genetically-Defined Ancestry Groups Rows represent ancestry groups based on genetic ancestry. Columns represent self-identified ancestry. Values represent the number of individuals in each ancestry group and how they self-identify.

	European	African	Latino
European	**5564**	**0**	**3**
African	**0**	**4532**	**1**

3.2.7 Supplementary Figure

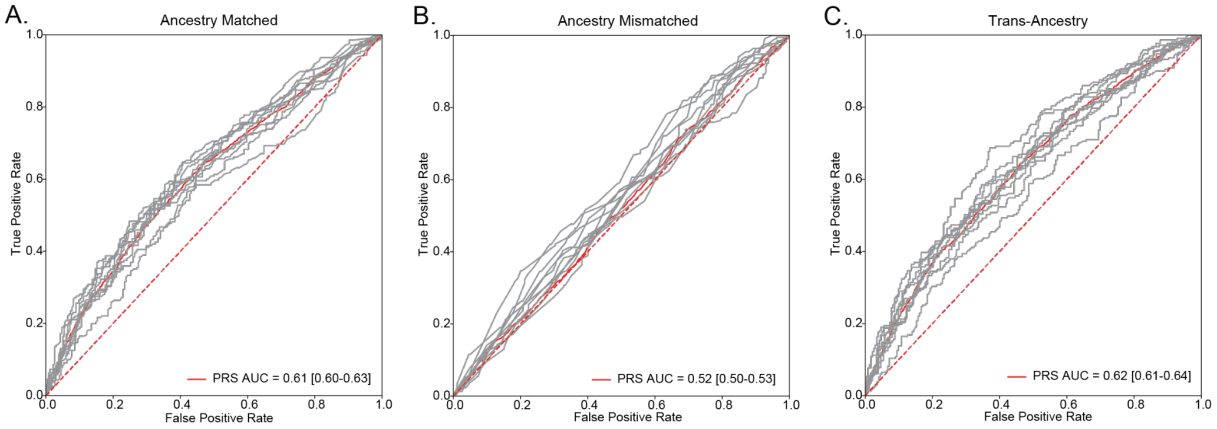

Figure S3.6 Performance of Polygenic Risk Scores Constructed from Different Ancestral Backgrounds in African Ancestry Group in ELLIPSE Consortium ROC curve for genetic prediction of prostate cancer risk in ELLIPSE Consortium African ancestry group (n=4,533) using: **(A)** Polygenic risk scores constructed from 10 African ancestry-specific variants only. **(B)** Polygenic risk scores constructed from 7 European ancestry-specific variants only. **(C)** Polygenic risk scores constructed from 14 trans-ancestry specific variants only.

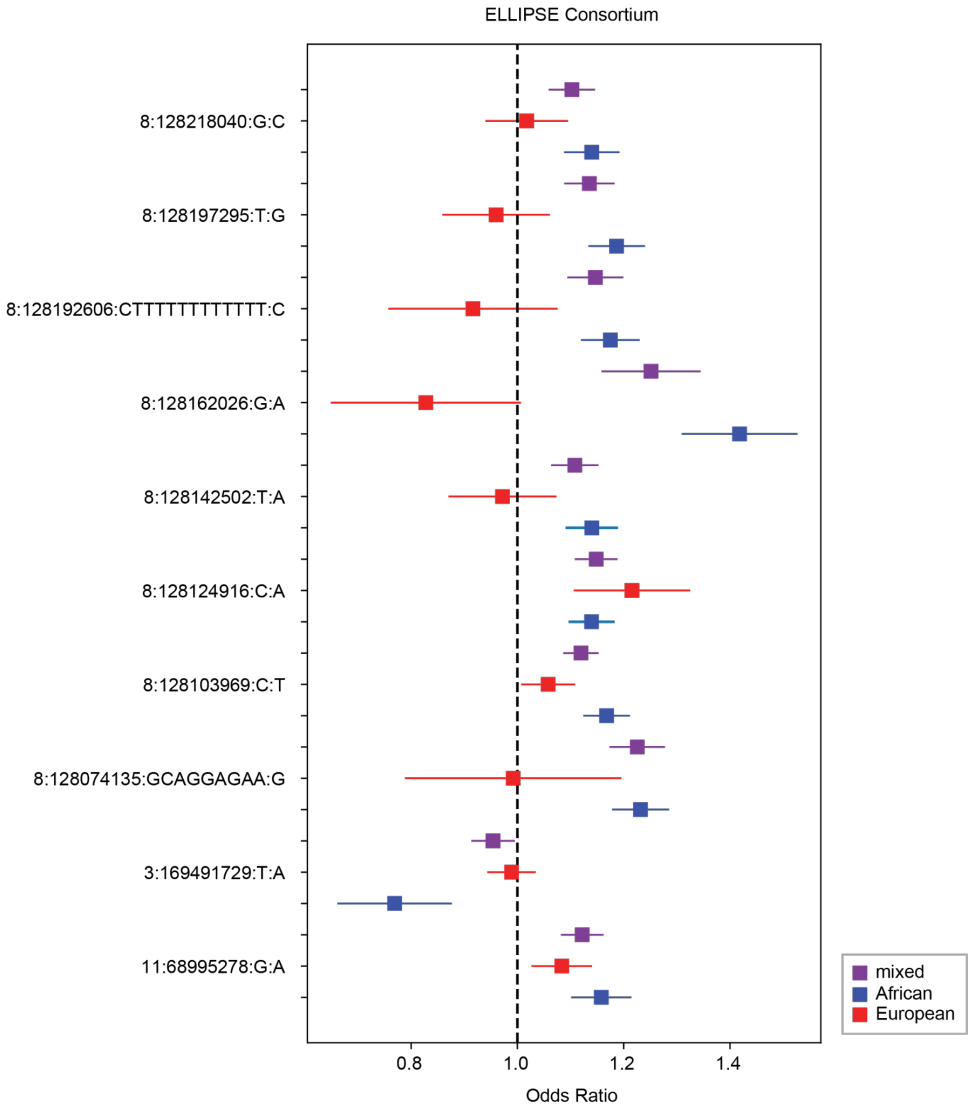

Figure S3.7 Comparison of African Prostate Cancer Variant Odds Ratios in European, African, and Trans-Ancestry Groups in ELLIPSE Consortium Odds ratios for 10 African ancestry-specific PrCa variants for prostate cancer in the ELLIPSE Consortium for European (n=5,667), African (n=4,553) and Trans-Ancestry (n=10,100) ancestry groups.

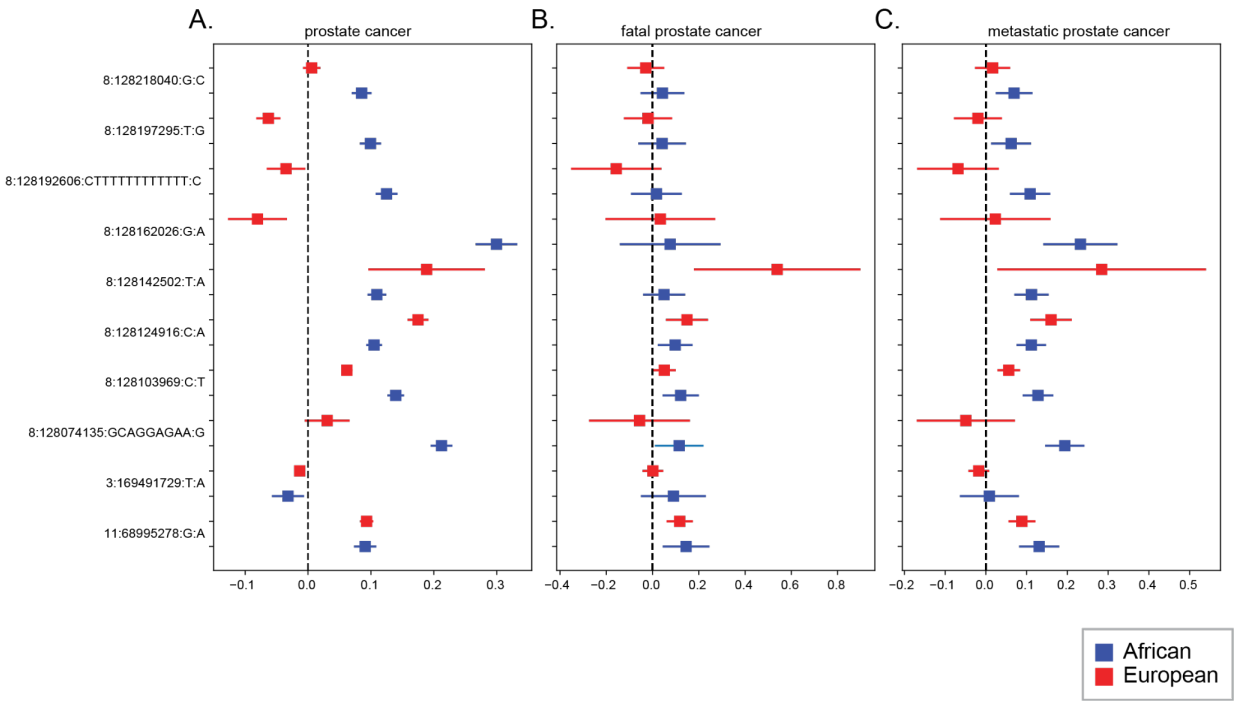

Figure S3.8 Comparison of African Prostate Cancer Variant Odds Ratios in European and African HARE Groups in Million Veteran Program Odds ratios for 10 African ancestry-specific PrCa variants for any prostate cancer **(A),** fatal prostate cancer **(B),** and metastatic prostate cancer **(C)** in the African (n=121,964) and European (n=461,627) HARE Million Veteran Program groups.

3.2.8 Author Contributions

Original concept: Stephen Ryan, Hannah Carter, Joshua Linscott, Meghana Pagadala

Genetic analysis: Meghana Pagadala

Validation Analysis: Julie Lynch, Tyler Seibert, Brent Rose, Richard Hauger, Matthew Panizzon, Meghana Pagadala, James Talwar

Manuscript Writing: Stephen Ryan, Meghana Pagadala, Hannah Carter, Moritz Hansen, Jesse Sammon, Matthew Hayn, Karim Kader

3.2.9 Acknowledgements

This research used data from the Million Veteran Program, Office of Research and Development, Veterans Health Administration. This research was supported by the Million Veteran Program MVP022 award #I01 CX001727 (PI: Richard L. Hauger MD). This publication does not represent the views of the Department of Veterans Affairs or the United States Government. This work was supported by National Institutes of Health grants #1F30CA247168, #T32CA067754 to Meghana S. Pagadala and an Emerging Leader Award from The Mark Foundation for Cancer Research grant #18-022-ELA to Hannah Carter.

The content of this chapter, in full, is a reprint of the following publication as it appears in medrxiv: Pagadala, Meghana; Linscott, Joshua; Talwar, James; Seibert, Tyler; Rose, Brent; Lynch, Julie; Panizzon, Matt; Hauger, Richard; Hansen, Moritz; Sammon, Jesse; Hayn, Matthew; Kader, Karim; Carter, Hannah; Ryan, Stephen. PRState: Incorporating Genetic Ancestry in Prostate Cancer Risk scores for African American Men. The dissertation author was the primary author of this material.

3.2.10 References

1. S. J. Henley, S. Jane Henley, E. M. Ward, S. Scott, J. Ma, R. N. Anderson, A. U. Firth, C. C. Thomas, F. Islami, H. K. Weir, D. R. Lewis, R. L. Sherman, M. Wu, V. B. Benard, L. C. Richardson, A. Jemal, K. Cronin, B. A. Kohler, Annual report to the nation on the status of cancer, part I: National cancer statistics. *Cancer*. **126** (2020), pp. 2225–2249.

2. O. W. Brawley, Prostate cancer epidemiology in the United States. *World Journal of Urology*. **30** (2012), pp. 195–200.

3. R. Siegel, J. Ma, Z. Zou, A. Jemal, Cancer statistics, 2014. *CA: A Cancer Journal for Clinicians*. **64** (2014), pp. 9–29.

4. L. N. Kolonel, B. E. Henderson, J. H. Hankin, A. M. Nomura, L. R. Wilkens, M. C. Pike, D. O. Stram, K. R. Monroe, M. E. Earle, F. S. Nagamine, A multiethnic cohort in Hawaii and Los Angeles: baseline characteristics. *Am. J. Epidemiol.* **151**, 346–357 (2000).

5. J. B. Hjelmborg, T. Scheike, K. Holst, A. Skytthe, K. L. Penney, R. E. Graff, E. Pukkala, K. Christensen, H.-O. Adami, N. V. Holm, E. Nuttall, S. Hansen, M. Hartman, K. Czene, J. R. Harris, J. Kaprio, L. A. Mucci, The heritability of prostate cancer in the Nordic Twin Study of Cancer. *Cancer Epidemiol. Biomarkers Prev.* **23**, 2303–2310 (2014).

6. P. Lichtenstein, N. V. Holm, P. K. Verkasalo, A. Iliadou, J. Kaprio, M. Koskenvuo, E. Pukkala, A. Skytthe, K. Hemminki, Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. *N. Engl. J. Med.* **343**, 78–85 (2000).

7. L. A. Mucci, J. B. Hjelmborg, J. R. Harris, K. Czene, D. J. Havelick, T. Scheike, R. E. Graff, K. Holst, S. Möller, R. H. Unger, C. McIntosh, E. Nuttall, I. Brandt, K. L. Penney, M. Hartman, P. Kraft, G. Parmigiani, K. Christensen, M. Koskenvuo, N. V. Holm, K. Heikkilä, E. Pukkala, A. Skytthe, H.-O. Adami, J. Kaprio, Nordic Twin Study of Cancer (NorTwinCan) Collaboration, Familial Risk and Heritability of Cancer Among Twins in Nordic Countries. *JAMA*. **315**, 68–76 (2016).

8. S. Benafif, Z. Kote-Jarai, R. A. Eeles, A Review of Prostate Cancer Genome-Wide Association Studies (GWAS). *Cancer Epidemiology Biomarkers & Prevention*. **27** (2018), pp. 845–857.

9. R. Eeles, C. Goh, E. Castro, E. Bancroft, M. Guy, A. A. Al Olama, D. Easton, Z. Kote-Jarai, The genetic epidemiology of prostate cancer and its clinical implications. *Nature Reviews Urology*. **11** (2014), pp. 18–31.

10. F. R. Schumacher, A. A. Al Olama, S. I. Berndt, S. Benlloch, M. Ahmed, E. J. Saunders, T. Dadaev, D. Leongamornlert, E. Anokian, C. Cieza-Borrella, C. Goh, M. N. Brook, X. Sheng, L. Fachal, J. Dennis, J. Tyrer, K. Muir, A. Lophatananon, V. L. Stevens, S. M. Gapstur, B. D. Carter, C. M. Tangen, P. J. Goodman, I. M. Thompson Jr, J. Batra, S. Chambers, L. Moya, J. Clements, L. Horvath, W. Tilley, G. P. Risbridger, H. Gronberg, M. Aly, T. Nordström, P.

Pharoah, N. Pashayan, J. Schleutker, T. L. J. Tammela, C. Sipeky, A. Auvinen, D. Albanes, S. Weinstein, A. Wolk, N. Håkansson, C. M. L. West, A. M. Dunning, N. Burnet, L. A. Mucci, E. Giovannucci, G. L. Andriole, O. Cussenot, G. Cancel-Tassin, S. Koutros, L. E. Beane Freeman, K. D. Sorensen, T. F. Orntoft, M. Borre, L. Maehle, E. M. Grindedal, D. E. Neal, J. L. Donovan, F. C. Hamdy, R. M. Martin, R. C. Travis, T. J. Key, R. J. Hamilton, N. E. Fleshner, A. Finelli, S. A. Ingles, M. C. Stern, B. S. Rosenstein, S. L. Kerns, H. Ostrer, Y.-J. Lu, H.-W. Zhang, N. Feng, X. Mao, X. Guo, G. Wang, Z. Sun, G. G. Giles, M. C. Southey, R. J. MacInnis, L. M. FitzGerald, A. S. Kibel, B. F. Drake, A. Vega, A. Gómez-Caamaño, R. Szulkin, M. Eklund, M. Kogevinas, J. Llorca, G. Castaño-Vinyals, K. L. Penney, M. Stampfer, J. Y. Park, T. A. Sellers, H.-Y. Lin, J. L. Stanford, C. Cybulski, D. Wokolorczyk, J. Lubinski, E. A. Ostrander, M. S. Geybels, B. G. Nordestgaard, S. F. Nielsen, M. Weischer, R. Bisbjerg, M. A. Røder, P. Iversen, H. Brenner, K. Cuk, B. Holleczek, C. Maier, M. Luedeke, T. Schnoeller, J. Kim, C. J. Logothetis, E. M. John, M. R. Teixeira, P. Paulo, M. Cardoso, S. L. Neuhausen, L. Steele, Y. C. Ding, K. De Ruyck, G. De Meerleer, P. Ost, A. Razack, J. Lim, S.-H. Teo, D. W. Lin, L. F. Newcomb, D. Lessel, M. Gamulin, T. Kulis, R. Kaneva, N. Usmani, S. Singhal, C. Slavov, V. Mitev, M. Parliament, F. Claessens, S. Joniau, T. Van den Broeck, S. Larkin, P. A. Townsend, C. Aukim-Hastie, M. Gago-Dominguez, J. E. Castelao, M. E. Martinez, M. J. Roobol, G. Jenster, R. H. N. van Schaik, F. Menegaux, T. Truong, Y. A. Koudou, J. Xu, K.-T. Khaw, L. Cannon-Albright, H. Pandha, A. Michael, S. N. Thibodeau, S. K. McDonnell, D. J. Schaid, S. Lindstrom, C. Turman, J. Ma, D. J. Hunter, E. Riboli, A. Siddiq, F. Canzian, L. N. Kolonel, L. Le Marchand, R. N. Hoover, M. J. Machiela, Z. Cui, P. Kraft, C. I. Amos, D. V. Conti, D. F. Easton, F. Wiklund, S. J. Chanock, B. E. Henderson, Z. Kote-Jarai, C. A. Haiman, R. A. Eeles, Profile Study, Australian Prostate Cancer BioResource (APCB), IMPACT Study, Canary PASS Investigators, Breast and Prostate Cancer Cohort Consortium (BPC3), PRACTICAL (Prostate Cancer Association Group to Investigate Cancer-Associated Alterations in the Genome) Consortium, Cancer of the Prostate in Sweden (CAPS), Prostate Cancer Genome-wide Association Study of Uncommon Susceptibility Loci (PEGASUS), Genetic Associations and Mechanisms in Oncology (GAME-ON)/Elucidating Loci Involved in Prostate Cancer Susceptibility (ELLIPSE) Consortium, Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. *Nat. Genet.* **50**, 928–936 (2018).

11. D. V. Conti, B. F. Darst, L. C. Moss, E. J. Saunders, X. Sheng, A. Chou, F. R. Schumacher, A. A. A. Olama, S. Benlloch, T. Dadaev, M. N. Brook, A. Sahimi, T. J. Hoffmann, A. Takahashi, K. Matsuda, Y. Momozawa, M. Fujita, K. Muir, A. Lophatananon, P. Wan, L. Le Marchand, L. R. Wilkens, V. L. Stevens, S. M. Gapstur, B. D. Carter, J. Schleutker, T. L. J. Tammela, C. Sipeky, A. Auvinen, G. G. Giles, M. C. Southey, R. J. MacInnis, C. Cybulski, D. Wokołorczyk, J. Lubiński, D. E. Neal, J. L. Donovan, F. C. Hamdy, R. M. Martin, B. G. Nordestgaard, S. F. Nielsen, M. Weischer, S. E. Bojesen, M. A. Røder, P. Iversen, J. Batra, S. Chambers, L. Moya, L. Horvath, J. A. Clements, W. Tilley, G. P. Risbridger, H. Gronberg, M. Aly, R. Szulkin, M. Eklund, T. Nordström, N. Pashayan, A. M. Dunning, M. Ghoussaini, R. C. Travis, T. J. Key, E. Riboli, J. Y. Park, T. A. Sellers, H.-Y. Lin, D. Albanes, S. J. Weinstein, L. A. Mucci, E. Giovannucci, S. Lindstrom, P. Kraft, D. J. Hunter, K. L. Penney, C. Turman, C. M. Tangen, P. J. Goodman, I. M. Thompson Jr, R. J. Hamilton, N. E. Fleshner, A. Finelli, M.-É. Parent, J. L. Stanford, E. A. Ostrander, M. S. Geybels, S. Koutros, L. E. B. Freeman, M. Stampfer, A. Wolk, N. Håkansson, G. L. Andriole, R. N. Hoover, M. J. Machiela, K. D. Sørensen, M. Borre, W. J. Blot, W. Zheng, E. D. Yeboah, J. E. Mensah, Y.-

J. Lu, H.-W. Zhang, N. Feng, X. Mao, Y. Wu, S.-C. Zhao, Z. Sun, S. N. Thibodeau, S. K. McDonnell, D. J. Schaid, C. M. L. West, N. Burnet, G. Barnett, C. Maier, T. Schnoeller, M. Luedeke, A. S. Kibel, B. F. Drake, O. Cussenot, G. Cancel-Tassin, F. Menegaux, T. Truong, Y. A. Koudou, E. M. John, E. M. Grindedal, L. Maehle, K.-T. Khaw, S. A. Ingles, M. C. Stern, A. Vega, A. Gómez-Caamaño, L. Fachal, B. S. Rosenstein, S. L. Kerns, H. Ostrer, M. R. Teixeira, P. Paulo, A. Brandão, S. Watya, A. Lubwama, J. T. Bensen, E. T. H. Fontham, J. Mohler, J. A. Taylor, M. Kogevinas, J. Llorca, G. Castaño-Vinyals, L. Cannon-Albright, C. C. Teerlink, C. D. Huff, S. S. Strom, L. Multigner, P. Blanchet, L. Brureau, R. Kaneva, C. Slavov, V. Mitev, R. J. Leach, B. Weaver, H. Brenner, K. Cuk, B. Holleczek, K.-U. Saum, E. A. Klein, A. W. Hsing, R. A. Kittles, A. B. Murphy, C. J. Logothetis, J. Kim, S. L. Neuhausen, L. Steele, Y. C. Ding, W. B. Isaacs, B. Nemesure, A. J. M. Hennis, J. Carpten, H. Pandha, A. Michael, K. De Ruyck, G. De Meerleer, P. Ost, J. Xu, A. Razack, J. Lim, S.-H. Teo, L. F. Newcomb, D. W. Lin, J. H. Fowke, C. Neslund-Dudas, B. A. Rybicki, M. Gamulin, D. Lessel, T. Kulis, N. Usmani, S. Singhal, M. Parliament, F. Claessens, S. Joniau, T. Van den Broeck, M. Gago-Dominguez, J. E. Castelao, M. E. Martinez, S. Larkin, P. A. Townsend, C. Aukim-Hastie, W. S. Bush, M. C. Aldrich, D. C. Crawford, S. Srivastava, J. C. Cullen, G. Petrovics, G. Casey, M. J. Roobol, G. Jenster, R. H. N. van Schaik, J. J. Hu, M. Sanderson, R. Varma, R. McKean-Cowdin, M. Torres, N. Mancuso, S. I. Berndt, S. K. Van Den Eeden, D. F. Easton, S. J. Chanock, M. B. Cook, F. Wiklund, H. Nakagawa, J. S. Witte, R. A. Eeles, Z. Kote-Jarai, C. A. Haiman, 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).

12. Z. Du, H. Hopp, S. A. Ingles, C. Huff, X. Sheng, B. Weaver, M. Stern, T. J. Hoffmann, E. M. John, S. K. Van Den Eeden, S. Strom, R. J. Leach, I. M. Thompson Jr, J. S. Witte, D. V. Conti, C. A. Haiman, A genome-wide association study of prostate cancer in Latinos. *Int. J. Cancer*. **146**, 1819–1826 (2020).

13. D. V. Conti, K. Wang, X. Sheng, J. T. Bensen, D. J. Hazelett, M. B. Cook, S. A. Ingles, R. A. Kittles, S. S. Strom, B. A. Rybicki, B. Nemesure, W. B. Isaacs, J. L. Stanford, W. Zheng, M. Sanderson, E. M. John, J. Y. Park, J. Xu, V. L. Stevens, S. I. Berndt, C. D. Huff, Z. Wang, E. D. Yeboah, Y. Tettey, R. B. Biritwum, A. A. Adjei, E. Tay, A. Truelove, S. Niwa, T. A. Sellers, K. Yamoah, A. B. Murphy, D. C. Crawford, S. M. Gapstur, W. S. Bush, M. C. Aldrich, O. Cussenot, G. Petrovics, J. Cullen, C. Neslund-Dudas, M. C. Stern, Z.-K. Jarai, K. Govindasami, A. P. Chokkalingam, A. W. Hsing, P. J. Goodman, T. Hoffmann, B. F. Drake, J. J. Hu, P. E. Clark, S. K. Van Den Eeden, P. Blanchet, J. H. Fowke, G. Casey, A. J. M. Hennis, Y. Han, A. Lubwama, I. M. Thompson Jr, R. Leach, D. F. Easton, F. Schumacher, D. J. Van den Berg, S. M. Gundell, A. Stram, P. Wan, L. Xia, L. C. Pooler, J. L. Mohler, E. T. H. Fontham, G. J. Smith, J. A. Taylor, S. Srivastava, R. A. Eeles, J. Carpten, A. S. Kibel, L. Multigner, M.-E. Parent, F. Menegaux, G. Cancel-Tassin, E. A. Klein, L. Brureau, D. O. Stram, S. Watya, S. J. Chanock, J. S. Witte, W. J. Blot, B. E. Henderson, C. A. Haiman, PRACTICAL/ELLIPSE Consortium, Two Novel Susceptibility Loci for Prostate Cancer in Men of African Ancestry. *J. Natl. Cancer Inst.* **109** (2017), doi:10.1093/jnci/djx084.

14. B. T. Helfand, J. Kearns, C. Conran, J. Xu, Clinical validity and utility of genetic risk scores in prostate cancer. *Asian J. Androl.* **18**, 509–514 (2016).

15. M. A. Liss, J. Xu, H. Chen, A. K. Kader, Prostate genetic score (PGS-33) is independently associated with risk of prostate cancer in the PLCO trial. *Prostate*. **75**, 1322–1328 (2015).

16. J. Sun, R. Na, F.-C. Hsu, S. L. Zheng, F. Wiklund, L. D. Condreay, J. M. Trent, J. Xu, Genetic score is an objective and better measurement of inherited risk of prostate cancer than family history. *Eur. Urol.* **63**, 585–587 (2013).

17. R. A. Karunamuni, M.-P. Huynh-Le, C. C. Fan, W. Thompson, R. A. Eeles, Z. Kote-Jarai, K. Muir, A. Lophatananon, UKGPCS collaborators, J. Schleutker, N. Pashayan, J. Batra, APCB BioResource (Australian Prostate Cancer BioResource), H. Grönberg, E. I. Walsh, E. L. Turner, A. Lane, R. M. Martin, D. E. Neal, J. L. Donovan, F. C. Hamdy, B. G. Nordestgaard, C. M. Tangen, R. J. MacInnis, A. Wolk, D. Albanes, C. A. Haiman, R. C. Travis, J. L. Stanford, L. A. Mucci, C. M. L. West, S. F. Nielsen, A. S. Kibel, F. Wiklund, O. Cussenot, S. I. Berndt, S. Koutros, K. D. Sørensen, C. Cybulski, E. M. Grindedal, J. Y. Park, S. A. Ingles, C. Maier, R. J. Hamilton, B. S. Rosenstein, A. Vega, IMPACT Study Steering Committee and Collaborators, M. Kogevinas, K. L. Penney, M. R. Teixeira, H. Brenner, E. M. John, R. Kaneva, C. J. Logothetis, S. L. Neuhausen, A. Razack, L. F. Newcomb, Canary PASS Investigators, M. Gamulin, N. Usmani, F. Claessens, M. Gago-Dominguez, P. A. Townsend, M. J. Roobol, W. Zheng, Profile Study Steering Committee, I. G. Mills, O. A. Andreassen, A. M. Dale, T. M. Seibert, PRACTICAL Consortium, Additional SNPs improve risk stratification of a polygenic hazard score for prostate cancer. *Prostate Cancer Prostatic Dis.* **24**, 532–541 (2021).

18. M.-P. Huynh-Le, R. Karunamuni, C. C. Fan, W. K. Thompson, K. Muir, A. Lophatananon, K. Tye, A. Wolk, N. Håkansson, I. G. Mills, O. A. Andreassen, A. M. Dale, T. M. Seibert, PRACTICAL Consortium, Common genetic and clinical risk factors: association with fatal prostate cancer in the Cohort of Swedish Men. *Prostate Cancer Prostatic Dis.* (2021), doi:10.1038/s41391-021-00341-4.

19. M.-P. Huynh-Le, C. C. Fan, R. Karunamuni, W. K. Thompson, M. E. Martinez, R. A. Eeles, Z. Kote-Jarai, K. Muir, J. Schleutker, N. Pashayan, J. Batra, H. Grönberg, D. E. Neal, J. L. Donovan, F. C. Hamdy, R. M. Martin, S. F. Nielsen, B. G. Nordestgaard, F. Wiklund, C. M. Tangen, G. G. Giles, A. Wolk, D. Albanes, R. C. Travis, W. J. Blot, W. Zheng, M. Sanderson, J. L. Stanford, L. A. Mucci, C. M. L. West, A. S. Kibel, O. Cussenot, S. I. Berndt, S. Koutros, K. D. Sørensen, C. Cybulski, E. M. Grindedal, F. Menegaux, K.-T. Khaw, J. Y. Park, S. A. Ingles, C. Maier, R. J. Hamilton, S. N. Thibodeau, B. S. Rosenstein, Y.-J. Lu, S. Watya, A. Vega, M. Kogevinas, K. L. Penney, C. Huff, M. R. Teixeira, L. Multigner, R. J. Leach, L. Cannon-Albright, H. Brenner, E. M. John, R. Kaneva, C. J. Logothetis, S. L. Neuhausen, K. De Ruyck, H. Pandha, A. Razack, L. F. Newcomb, J. H. Fowke, M. Gamulin, N. Usmani, F. Claessens, M. Gago-Dominguez, P. A. Townsend, W. S. Bush, M. J. Roobol, M.-É. Parent, J. J. Hu, I. G. Mills, O. A. Andreassen, A. M. Dale, T. M. Seibert, UKGPCS collaborators, APCB (Australian Prostate Cancer BioResource), NC-LA PCaP Investigators, IMPACT Study Steering Committee and Collaborators, Canary PASS Investigators, Profile Study Steering Committee, PRACTICAL Consortium, Polygenic hazard score is associated with prostate cancer in multi-ethnic populations. *Nat. Commun.* **12**, 1236 (2021).

20. T. M. Seibert, C. C. Fan, Y. Wang, V. Zuber, R. Karunamuni, J. K. Parsons, R. A. Eeles, D.

F. Easton, Z. Kote-Jarai, A. A. Al Olama, S. B. Garcia, K. Muir, H. Grönberg, F. Wiklund, M. Aly, J. Schleutker, C. Sipeky, T. L. Tammela, B. G. Nordestgaard, S. F. Nielsen, M. Weischer, R. Bisbjerg, M. A. Røder, P. Iversen, T. J. Key, R. C. Travis, D. E. Neal, J. L. Donovan, F. C. Hamdy, P. Pharoah, N. Pashayan, K.-T. Khaw, C. Maier, W. Vogel, M. Luedeke, K. Herkommer, A. S. Kibel, C. Cybulski, D. Wokolorczyk, W. Kluzniak, L. Cannon-Albright, H. Brenner, K. Cuk, K.-U. Saum, J. Y. Park, T. A. Sellers, C. Slavov, R. Kaneva, V. Mitev, J. Batra, J. A. Clements, A. Spurdle, M. R. Teixeira, P. Paulo, S. Maia, H. Pandha, A. Michael, A. Kierzek, D. S. Karow, I. G. Mills, O. A. Andreassen, A. M. Dale, PRACTICAL Consortium*, Polygenic hazard score to guide screening for aggressive prostate cancer: development and validation in large scale cohorts. *BMJ*. **360**, j5757 (2018).

21. S. Das, L. Forer, S. Schönherr, C. Sidore, A. E. Locke, A. Kwong, S. I. Vrieze, E. Y. Chew, S. Levy, M. McGue, D. Schlessinger, D. Stambolian, P.-R. Loh, W. G. Iacono, A. Swaroop, L. J. Scott, F. Cucca, F. Kronenberg, M. Boehnke, G. R. Abecasis, C. Fuchsberger, Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).

22. K. K. Kidd, U. Soundararajan, H. Rajeevan, A. J. Pakstis, K. N. Moore, J. D. Ropero-Miller, The redesigned Forensic Research/Reference on Genetics-knowledge base, FROG-kb. *Forensic Sci. Int. Genet.* **33**, 33–37 (2018).

23. S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, P. C. Sham, PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*. **81** (2007), pp. 559–575.

24. H. Fang, Q. Hui, J. Lynch, J. Honerlaw, T. L. Assimes, J. Huang, M. Vujkovic, S. M. Damrauer, S. Pyarajan, J. M. Gaziano, S. L. DuVall, C. J. O'Donnell, K. Cho, K.-M. Chang, P. W. F. Wilson, P. S. Tsao, VA Million Veteran Program, Y. V. Sun, H. Tang, Harmonizing genetic ancestry and self-identified race/ethnicity in genome-wide association studies. *Am. J. Hum. Genet.* **105**, 763–772 (2019).

25. K. K. Kidd, W. C. Speed, A. J. Pakstis, M. R. Furtado, R. Fang, A. Madbouly, M. Maiers, M. Middha, F. R. Friedlaender, J. R. Kidd, Progress toward an efficient panel of SNPs for ancestry inference. *Forensic Sci. Int. Genet.* **10**, 23–32 (2014).

26. C. A. Haiman, N. Patterson, M. L. Freedman, S. R. Myers, M. C. Pike, A. Waliszewska, J. Neubauer, A. Tandon, C. Schirmer, G. J. McDonald, S. C. Greenway, D. O. Stram, L. Le Marchand, L. N. Kolonel, M. Frasco, D. Wong, L. C. Pooler, K. Ardlie, I. Oakley-Girvan, A. S. Whittemore, K. A. Cooney, E. M. John, S. A. Ingles, D. Altshuler, B. E. Henderson, D. Reich, Multiple regions within 8q24 independently affect risk for prostate cancer. *Nat. Genet.* **39**, 638–644 (2007).

27. I. Cheng, S. J. Plummer, E. Jorgenson, X. Liu, B. A. Rybicki, G. Casey, J. S. Witte, 8q24 and prostate cancer: association with advanced disease and meta-analysis. *Eur. J. Hum. Genet.* **16**, 496–505 (2008).

3.3.1 Abstract

Genetic scores may provide an objective measure of a man's risk of prostate cancer and thus inform screening decisions. We evaluated whether a polygenic hazard score based on 290 genetic variants (PHS290) is associated with risk of prostate cancer in a diverse population, including Black men, who have higher average risk of prostate cancer death but are often treated as a homogeneous, high-risk group. This was a retrospective analysis of Million Veteran Program (MVP), a national, population-based cohort study of United States military veterans conducted 2011-2021. Cox proportional hazards analyses tested for association of genetic and other risk factors (including self-reported race/ethnicity and family history) with age at death from prostate cancer, age at diagnosis of metastatic (nodal or distant) prostate cancer, and age at diagnosis of any prostate cancer. 590,750 male participants were included. Median age at last follow-up was 69 years. PHS290 was associated with fatal prostate cancer in the full cohort and for each racial/ethnic group ($p<10^{-10}$). Comparing men in the highest 20% of PHS290 to those in the lowest 20%, the hazard ratio for fatal prostate cancer was 4.42 [95% CI: 3.91-5.02]. When accounting for guideline-recommended risk factors (family history, race/ethnicity), PHS290 remained the strongest independent predictor of any, metastatic, and fatal prostate cancer. PHS290 stratified US veterans of diverse ancestry for lifetime risk of prostate cancer, including metastatic and fatal cancer. Predicting genetic risk of lethal prostate cancer with PHS290 might inform individualized decisions about prostate cancer screening.

3.3.2 Introduction

Prostate cancer is the most diagnosed and second deadliest cancer in men(*1*). Despite the enormous mortality from this disease, early detection of prostate cancer remains controversial.

Screening all men via prostate-specific antigen (PSA) testing, regardless of underlying risk, has been shown to reduce prostate cancer deaths by 27% but also results in numerous false positive results and frequent overdiagnosis of indolent prostate cancer that may never have become symptomatic(*2–4*). These overdiagnoses often lead to unnecessary treatment, with attendant side effects and societal costs. A better strategy is to target PSA screening to those men at higher risk of developing metastatic or fatal prostate cancer.

As one of the most heritable cancers(*5*), genetic risk stratification is a promising approach for identifying individuals at higher risk of developing metastatic or fatal prostate cancer(*1*, *3*, *6*). Measures of genetic risk have proven highly effective for predicting lifetime risk of being diagnosed with prostate cancer, outperforming family history or other clinical risk factors(*7–10*). Rather than only predicting lifetime risk, however, an ideal genetic test would focus on clinically significant prostate cancer and estimate age-specific risk. Prostate cancer is highly age dependent, with very low incidence before 50 years of age and increasing exponentially as men get older(*11*, *12*). Absolute incidence of *aggressive* prostate cancer also increases with age(*11*, *12*). Meanwhile, some men with high genetic risk develop aggressive prostate cancer at a younger age and are at particular risk of dying from this disease. Age-specific genetic risk could inform individualized decisions about PSA testing, in the context of a given man's overall health and competing causes of mortality.

A major limitation of early studies of polygenic risk was an exclusive focus on men of European ancestry(*13*, *14*). Such systematic bias may exacerbate existing health disparities in prostate cancer incidence and health outcomes(*15*, *16*). This is particularly worrisome for men of African ancestry, who have a higher overall incidence of metastatic and fatal prostate cancer than

men of European or Asian ancestry(*17*, *18*). Recent efforts have incorporated data from more diverse populations, yielding improved performance in these groups(*19–23*).

Our group has developed a risk prediction tool called a polygenic hazard score (PHS) that identifies men who are likely to develop clinically significant prostate cancers at younger ages. This score, which can be calculated from a single saliva sample at any point in a man's life, was strongly associated with age at diagnosis of clinically significant prostate cancer in large datasets(*10*, *19*, *22*). The score also improved the accuracy of conventional screening with PSA(*7*, *10*, *12*). We subsequently expanded the model to optimize performance in men of all ancestries, particularly men with African ancestry(*19*, *22*, *24*). Here, we seek to validate the ability of the PHS to identify men at risk of metastatic or fatal prostate cancer within the Million Veteran Program (MVP) longitudinal cohort, one of the largest and most racially and ethnically diverse populations studied to date(*25*).

3.3.3 Results

3.3.3.1 PHS290 Score

The distribution of PHS290 in Non-Hispanic Whites was similar to that reported previously for men of European ancestry (mean=9.37, SD=0.37)(*24*). Mean PHS290 did vary by self-reported race/ethnicity with statistically significant differences between all groups (ANOVA $p < 10^{-16}$; all pair-wise *t*-tests $p < 10^{-14}$). The distribution for the Hispanic ancestry group overlapped closely with that of the European group (mean=9.35, SD=0.37), while PHS290 tended to be lower among Asian men (mean=9.18, SD=0.35) and higher among Black men (mean=9.56, SD=0.34, t-test) (**Figure 3.13, Figure S3.9**).

3.3.3.2 Association of PHS290 with Prostate Cancer

PHS290 was associated with age at diagnosis of prostate cancer, with age at development of prostate cancer nodal/distant metastases, and with age at death from prostate cancer **(Table 3.4)**. These associations also held in all racial/ethnic subgroup analyses with >100 events. Comparing 80^{th} and 20^{th} percentiles of genetic risk in the full dataset, men with higher PHS290 had an $HR_{80/20}$ of 5.20 [95% CI: 5.09-5.31] for any prostate cancer, $HR_{80/20}$ of 4.89 [95% CI: 4.57-5.21] for metastatic prostate cancer, and $HR_{80/20}$ of 4.42 [95% CI: 3.91-5.02] for fatal prostate cancer. Cause-specific cumulative incidence curves for various PHS290 percentile groups demonstrated risk stratification **(Figure 3.14)**. Consistent with prior reports, Black men had a higher average incidence of prostate cancer than Non-Hispanic White men. However, the incidence among Black men with low PHS290 was comparable to that of the average among Non-Hispanic White men **(Figure 3.14)**.

3.3.3.3 Race/Ethnicity, Family History, and PHS290

Race/ethnicity and family history were each associated with every clinical endpoint in univariable models. Men with a family history of prostate cancer had a HR of 1.83 [1.53-2.17] for dying of prostate cancer. The race/ethnicity associations were largely driven by an increased risk for Black men. Compared to Non-Hispanic White men, Black men had a HR of 2.53 [2.14-2.92] for dying of prostate cancer.

PHS290 remained an independent predictor of prostate cancer risk—including prostate cancer death—when accounting for race/ethnicity and family history (**Table 3.5**). The multivariable model improved prediction for each clinical endpoint over the common risk factors alone (ANOVA $p < 2.2 \times 10^{-16}$). Independent of ancestry and family history, a high PHS290 (top

20%) approximately quadrupled a man's risk of death from prostate cancer, compared to a low PHS290 (bottom 20%) **(Table 3.6)**.

3.3.3.4 Genetic Ancestry and Genetic Principal Components

Results of the above analyses using genetic ancestry and the top 10 genetic principal components are reported in the Supplementary Material. PHS290 was significantly and independently associated with all three prostate cancer endpoints in each case.

3.3.4 Discussion

PHS290 was associated with lifetime prostate-cancer-specific mortality in this large and diverse dataset. Even when accounting for current guideline-recommended risk factors (family history and race/ethnicity), PHS290 remained a strong independent predictor of dying from prostate cancer. The genomic score was also associated with age at diagnosis of metastasis from prostate cancer (nodal or distant) and with age at diagnosis of any prostate cancer. Metastatic prostate cancer has poor prognostic outcomes and is a major driver of pain, disability and aggressive medical therapy(*32*). To our knowledge, this study is the first to show the association of a genomic score with lifetime risk of metastatic prostate cancer. This study also represents the largest and most racially/ethnically diverse independent validation of the association of a polygenic score with lifetime risk of fatal prostate cancer.

Black men in the US are substantially more likely to develop metastatic disease and to die from prostate cancer(*18*). The causes of this disparity are likely a combination of genetic, environmental, and social factors, including systemic racism(*33–37*). National guidelines recommend consideration of prostate screening in men of African ancestry at a younger age and

that screening occur at more frequent intervals(*38*). The results of the present study confirm a generally great risk of prostate cancer among Black men but also demonstrate that Black men have variable levels of lifetime risk and should not be treated as a homogeneous group. Black men with low PHS290 had a prostate cancer risk comparable to average-risk Non-Hispanic White men, while Black men with high PHS290 had the highest risk of all subgroups. PHS290 can identify those more likely to develop lethal prostate cancer and may facilitate personalized screening recommendations.

Intriguingly, typical PHS290 scores differed between racial/ethnic groups, with the mean PHS290 slightly higher among Black men and slightly lower among East Asian men, compared to Hispanic and Non-Hispanic White men. These shifts in PHS290 distribution are consistent with reported differences in prostate cancer incidence across racial groups(*39–43*) and with a previous polygenic risk meta-analysis(*21*). Higher overall PHS290 scores in African ancestry group may point to true differences in prostate cancer risk but could also be inflated by minor allele frequency (MAF) differences between ancestry groups. Incorporating approaches for local ancestry and admixture can also boost genetic model performance and should be explored further to improve the predictive accuracy of polygenic scores(*44*).

Family history is another important clinical consideration in prostate cancer screening decisions(*38*, *45–48*). Prior studies have found polygenic scores to be the most important risk factor for prostate cancer (in men without known rare pathogenic mutations), with family history typically also independently predictive in multivariable models, possibly by capturing yet unknown genetic factors and/or shared familial environmental factors(*7*, *9*, *22*, *49*, *50*). Among MVP participants, family history of prostate cancer was independently associated with prostate cancer risk in a multivariable model that included race/ethnicity and PHS290. The relationship of

environmental exposures, family history, and prostate cancer risk merit further investigation(*9*), particularly in groups like veterans who may have been exposed to rare carcinogens(*51*).

The present study builds on prior work that reported the performance of polygenic scores in non-Europeans(*20–22*, *24*, *52*) and is consistent with those prior studies in showing a strong association of polygenic scores with prostate cancer risk, including death from prostate cancer(*9*, *22*, *49*). Polygenic hazard scores designed to incorporate the strong age-dependence of prostate cancer have also been shown to increase the accuracy of conventional prostate cancer screening(*7*, *10*, *12*). Population-level analyses of benefit, harm, and cost-effectiveness support incorporation of genomic risk into screening(*3*, *6*). The present study adds to the literature an independent validation in a dataset of over half a million men with diverse race/ethnicity and ancestry. Current clinical guidelines try to achieve targeted, or risk-stratified, screening by recommending each man discuss his individual risk factors, emphasizing race/ethnicity(*38*, *46–48*). It is particularly important, therefore, that this study was able to combine race/ethnicity and genetic risk to estimate the relative impact of each and to demonstrate that a polygenic score adds considerable information beyond race/ethnicity alone for a man's individual risk of metastasis or death from prostate cancer.

While PHS290 performed well in the present study to stratify men by genetic prostate cancer risk, the effect sizes estimated here are lower than those reported in previous studies. Smaller effect sizes were also seen when comparing the strength of association with age at diagnosis of prostate cancer within ancestry groups(*24*). Most likely, the discrepancy arises in differences in the populations studied; for example, the MVP dataset comes exclusively from a population of US veterans, with many receiving healthcare in a single US-based system, whereas the prior study used data from multiple countries and widely varying recruitment strategies. Patterns of screening, detection, and treatment of prostate cancer in the present dataset could be

different from clinical trial and case-control datasets used in previous work. Some of the difference in performance could also be explained by the fact that the testing datasets in the prior report for PHS290 had been included in the discovery of a majority of the 290 variants in the model; on the other hand, the testing datasets represented a very small proportion of the discovery datasets, and the model weights were estimated in an independent training dataset.

The present work shows that adding PHS290 to guideline-recommended risk factors improves risk stratification for meaningful clinical endpoints of death or metastases from prostate cancer. Men at highest risk of metastatic or fatal prostate cancer are potentially those most likely to benefit from screening. Prior studies have further suggested genetic scores could also add value even after results from screening or diagnostic tests are already available, but this needs further investigation(*53*, *54*). For example, one early detection strategy with strong evidence is early baseline PSA (e.g., at age 45-49)(*55*, *56*). This strategy has not yet been widely adopted in the U.S.(*57*), but future studies should evaluate whether PHS290 adds value in men where early baseline PSA is known and whether PHS290, if known prior to PSA testing, should inform the decision of whether to obtain an early baseline PSA test. Another compelling avenue for future studies is whether high genetic risk can be mitigated by lifestyle or other preventive intervention(*58*, *59*).

Notwithstanding the possible advantages to risk-based screening, men should still be cautioned on the risks of screening: false positive PSA testing results and overdiagnosis of non-threatening cancer. While men with high PHS290 have a higher absolute incidence of prostate cancer metastasis or death, they also remain at risk of developing low-grade cancers likely to be detected with screening. Strategies to mitigate screening harms should be appropriately applied,

including multiparametric MRI prior to biopsy and active surveillance for cancer diagnoses with favorable prognosis(*38*).

Limitations of this study include heterogeneity of prostate cancer screening and diagnostic pathways by clinicians across VA and other hospitals in the US that could potentially introduce noise, although this heterogeneity likely leads to underestimation of associations with prostate cancer. The natural language processing tool used to identify men with metastatic disease does not reliably distinguish regional nodal from distant metastases, so these were all considered as one endpoint. Finally, we acknowledge that while we have used race/ethnicity, genetic ancestry, and principal genetic components, none of these groups can account for—much less, disentangle—the complex web of biological and social factors associated with these categories. Further work will attempt to incorporate agnostic genetic ancestry groups and address impacts of admixture and local/regional genetic ancestry on risk stratification with PHS(*22*).

We show that PHS290 stratified US men for lifetime risk of any, metastatic, and fatal prostate cancer. Critically, this genetic risk stratification was successful within racial/ethnic subgroups in this diverse dataset. PHS290 was higher, on average, among Black men, who were also at higher risk from prostate cancer. The combination of race/ethnicity, family history, and PHS290 performed better than any single risk factor in identifying men at highest risk of prostate cancer metastasis and death. Predicting genetic risk of lethal prostate cancer with PHS290 might inform individualized decisions about screening and early cancer detection.

3.3.5 Methods

3.3.5.1 Participants

We retrospectively obtained data from the MVP, composed of individuals between ages 19 to over 100 years who were recruited from 63 Veterans Affairs Medical Centers across the United States (US). Recruitment for the MVP started in 2011, and all veterans were eligible for participation. Consent to participate and permission to re-contact was provided after counseling by research staff and mailing of informational materials. Study participation included consenting to access the participant's electronic health records for research purposes. The MVP received ethical and study protocol approval from the VA Central Institutional Review Board in accordance with the principles outlined in the Declaration of Helsinki.

Only men were included in this prostate cancer study, comprising 590,750 individuals of diverse self-reported race/ethnicity, over 100,000 of whom were Black or African American **(Table 3.1).** There were no inclusion or exclusion criteria for age. Median age at last follow-up was 69 years (interquartile range 59-74 years). Men not meeting the endpoint for each analysis were censored at age at last follow-up. Clinical information used for analyses was retrieved as described below in the Clinical Data Extraction section.

3.3.5.2 Genotype Data

All study participants provided blood samples for DNA extraction and genotyping. Researchers are provided data that is de-identified except for dates. Blood samples were collected by phlebotomists and banked at the VA Central Biorepository in Boston, MA, where DNA was extracted and shipped to two external centers for genotyping. DNA extracted from buffy coat was genotyped using a custom Affymetrix Axiom biobank array. The MVP 1.0 genotyping array contains a total of 723,305 variants, enriched for low frequency variants in African and Hispanic

populations and variants associated with diseases common to the VA population(*25*). The details on the quality control and imputation have been described previously(*26*).

3.3.5.3 Clinical Data Extraction

Each participant's electronic health record is integrated into the MVP biorepository. These records include International Classification of Diseases (ICD) diagnosis codes (ICD-9-CM and ICD-10-CM), procedure codes (ICD, Current Procedural Terminology, and Healthcare Common Procedure Coding (HCPCS)), laboratory values, medications, and clinical notes documenting VA care (inpatient and outpatient) and non-VA care paid for by the VA.

Prostate cancer diagnosis, age at diagnosis, and date of last follow-up were retrieved from the VA Corporate Data Warehouse based on ICD codes and VA Central Cancer Registry data. Age at diagnosis of metastasis (nodal and/or distant) was determined via a validated natural language processing tool and a search of individual participant's medical records in the Veterans Affairs system, as described previously(*27*). This tool was developed using data from over 1 million VA patients with prostate cancer; compared to manual chart review, the natural language processing tool had 92% sensitivity and 98% specificity for diagnosis of metastatic prostate cancer. Cause and date of death was collected from National Death Index. Participants with ICD10 code "C61" as underlying cause of death were considered to have died from prostate cancer. Age of death was determined from difference between year of death and year of birth.

3.3.5.4 Polygenic Hazard Score (PHS290)

The most recent version of the PHS, called PHS290, was calculated as the vector product of participants' genotype dosage (X_i) for 290 variants and the corresponding parameter estimates (β_i) from Cox proportional hazards regression:

$$PHS = \sum_{i}^{n} X_i \beta_i$$

The development of this score has been described elsewhere(*24*). Briefly, previously identified common variants associated with prostate cancer risk were simultaneously evaluated using a machine-learning least absolute shrinkage and selection operator (LASSO) approach to generate an optimal combined model for association with age at prostate cancer diagnosis.

We calculated PHS290 for each MVP participant. Distributions were visualized using histograms for each ancestry group. Differences in mean PHS290 between ancestry groups were assessed via ANOVA. In all statistical analyses, significance for association with clinical endpoints was set at a two-tailed alpha of 0.01. As in prior studies, *p*-values less than 10^{-16} were truncated at this value, as comparison of miniscule values is not likely to be meaningful(*7*, *10*, *12*, *19*). Subgroup analyses with less than 100 events are reported in the Supplemental Material(*28*).

3.3.5.5 Cox Proportional Hazards Analysis

We evaluated association of PHS290 with age at diagnosis of prostate cancer and with two important clinical endpoints extracted from clinical data: age at diagnosis of nodal and/or distant metastases from prostate cancer and age at death from prostate cancer (i.e., lifetime prostate-cancer-specific mortality). To visualize the association in the full dataset, we generated cause-specific cumulative incidence curves for each endpoint and each of several PHS290 risk

groups. Cox proportional hazards models were used to assess these associations in the full dataset and in each racial/ethnic group. Individuals not meeting the endpoint of interest were censored at age at last follow-up.

Effect sizes were estimated using hazard ratios (HRs) between risk strata, as described previously(*7*, *9*, *10*, *12*, *19*, *20*, *22*, *23*) and with previously defined thresholds for PHS290: 9.004659 (20^{th} quantile), 9.123500 (30^{th} quantile), 9.519703 (70^{th} quantile), 9.639068 (80^{th} quantile), 9.946332 (95^{th} quantile)(*24*). HRs for each ancestry group were calculated to make the following comparisons: $HR_{80/20}$, men in the highest 20% vs. lowest 20%; $HR_{95/50}$, men in the highest 5% of genetic risk vs. those with average risk (30–70th percentile); and $HR_{20/50}$, men in the lowest 20% vs. those with average risk.

3.3.5.6 Race/Ethnicity, Family History, and PHS290

To assess the independent predictive value of PHS290 beyond commonly used clinical risk factors, we tested a multivariable Cox proportional hazards model with self-reported race/ethnicity, family history, and PHS290(*7*, *9*, *22*). Family history was recorded as either the presence or absence of (one or more) first-degree relatives with prostate cancer. Cox proportional hazards models tested associations with any, metastatic, or fatal prostate cancer. For PHS290, the effect size was illustrated via the hazard ratio for the highest 20% vs. lowest 20% of genetic risk. Hazard ratios for racial/ethnic groups were estimated using Non-Hispanic White as the reference.

A univariable Cox proportional hazards model was applied to test for association of race/ethnicity with prostate cancer endpoints. Similarly, a univariable model tested for association of family history alone. The @anova function from the *R* 'survival' package (version 3.2-13; Therneau 2021) was used to compare the nested Cox models (multivariable vs. univariable), based

on the log partial likelihood of the model fits. Significance was set at a two-tailed alpha of 0.01 for the test of whether the multivariable model performed better than either univariable model alone.

3.3.5.7 Genetic Ancestry and Genetic Principal Components

MVP participants have been assigned genetic ancestry groups based on previous analyses(*29*). Briefly, a reference panel of 1000 Genomes Project(*30*) and Human Genome Diversity Project(*31*) individuals was constructed for preliminary PCA analysis. PC loadings of reference panel were projected onto PC loadings of MVP participant PC loadings, and assignments were made based on a random forest classifier. We repeated the prostate cancer analyses described above using genetic ancestry groups instead of self-reported race/ethnicity. We also repeated the above analyses when including the top 10 genetic principal components.

3.3.6 Figures

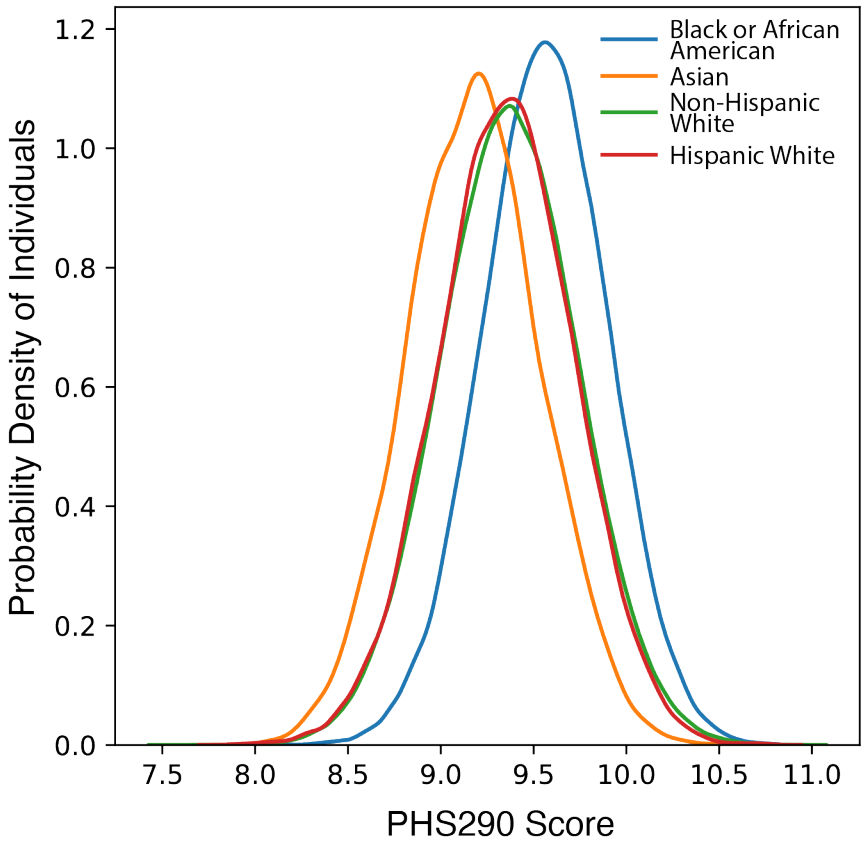

Figure 3.13 PHS290 Score Density Plots in Million Veteran Program PHS290 score density plot in select self-reported race/ethnicity groups.

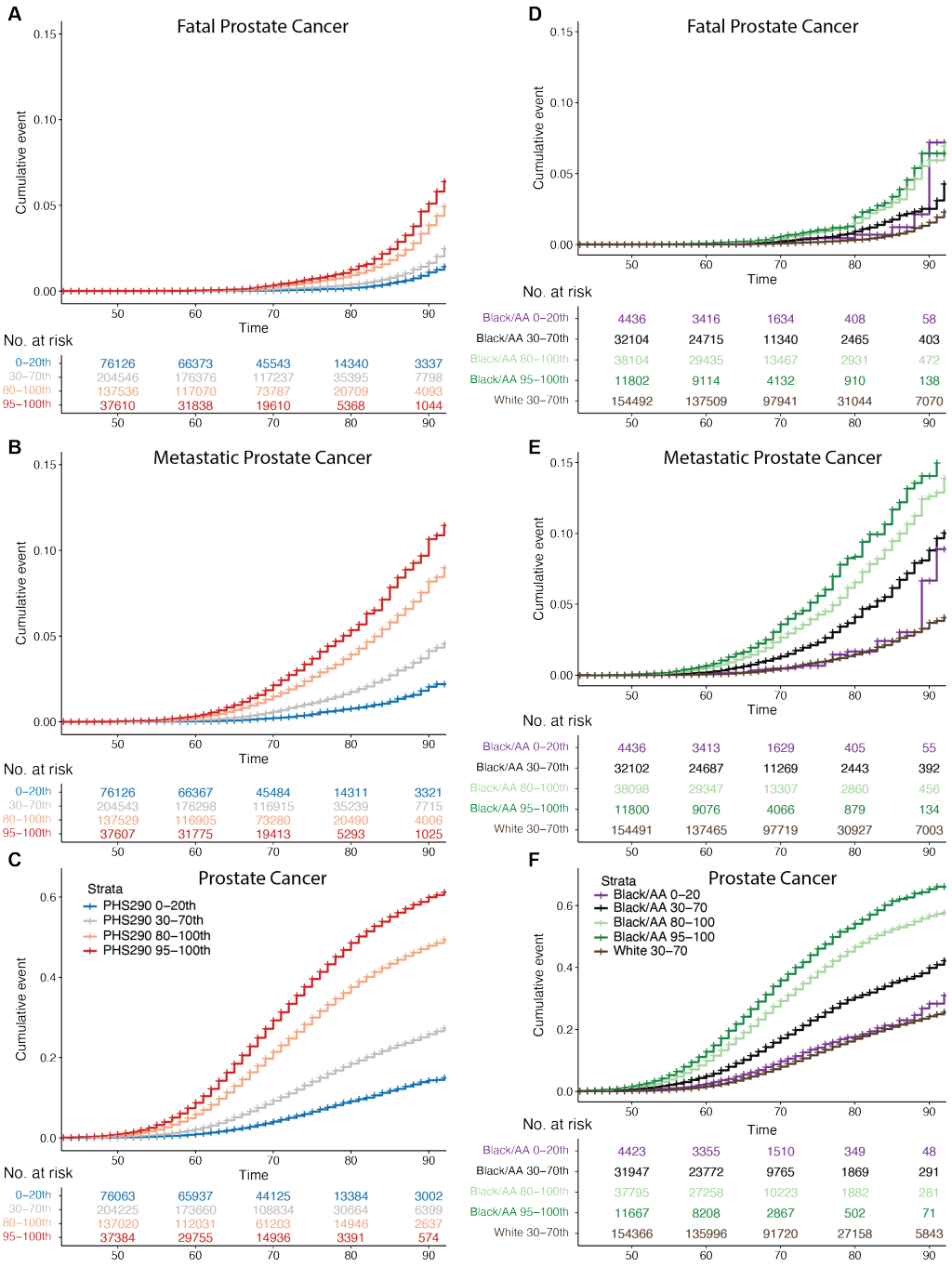

Figure 3.14 Million Veteran Program (MVP) Cause-specific Cumulative Incidence Cause-specific cumulative incidence within MVP, stratified by PHS290, for **(A)** fatal prostate cancer, **(B)** metastatic prostate cancer, and **(C)** prostate cancer. PHS290 percentile groups shown for each endpoint: 0-20th, 30-70th, 80-100th, and 95-100th. Cumulative incidence for Black or African American (Black/AA) men in several PHS290 strata compared with average-risk Non-Hispanic White (PHS290 30-70th percentiles) for **(D)** fatal prostate cancer, **(E)** metastatic prostate cancer, and **(F)** prostate cancer. The *y*-axis scale was adjusted for (C) and (E) to show the higher incidence values for any prostate cancer.

3.3.7 Tables

Table 3.4 Participant Characteristics for Self-Reported Race/Ethnicity Groups, n=590,750
Numbers in parentheses indicate participants with family history information also.

	All	Non-Hispanic White	Black or African American	Hispanic White	Asian	Native American	Pacific Islander	Other	Unknown
Participants									
All participants	590,750 (378,366)	420,473 (299,266)	102,203 (47,838)	27,651 (13,407)	6,644 (3,876)	5,835 (3,927)	3,246 (740)	8,226 (8,224)	16,472 (1,088)
Prostate cancer	69,137 (49,400)	48,339 (37,922)	15,748 (8,613)	2,120 (1,325)	376 (287)	504 (378)	236 (60)	720 (720)	1,094 (95)
Metastases from prostate cancer	6,413 (4,274)	4,299 (3,199)	1,604 (804)	213 (126)	33 (24)	48 (35)	18 (5)	68 (68)	130 (13)
Death from prostate cancer	1,858 (1,314)	1,354 (1,059)	363 (192)	59 (30)	5 (4)	10 (6)	4 (2)	19 (19)	44 (2)
Age demographics									
Age at diagnosis, median and IQR	67 [62-72]	68 [63-73]	63 [58-68]	66 [60-71]	67 [61-74]	65 [61-69]	65 [60-70]	64 [60-69]	67 [61-73]
Age at last follow-up, median and IQR	66 [59-74]	68 [62-75]	62 [55-70]	59 [47-71]	56 [41-70]	63 [56-72]	59 [49-70]	63 [57-71]	56 [42-70]

Table 3.5 Association of PHS290 with Any, Metastatic and Fatal Prostate Cancer Cox Proportional Hazards model results from association with age at prostate cancer, metastatic prostate cancer (nodal or distant) and death from prostate cancer. *P*-values reported are from univariable models using PHS290 as the sole predictor variable. Hazard ratios (HRs) compare men in various percentiles of genetic risk. $HR_{80/20}$: highest 20% ($\geq 80^{th}$ percentile of PHS290, using previously published thresholds for men <70 years old and no diagnosis of cancer) vs. average risk ($30\text{-}70^{th}$ percentile). $HR_{20/50}$: lowest 20% ($\leq 20^{th}$ percentile) vs. average risk. $HR_{80/50}$: highest 20% vs. average risk. $HR_{95/50}$: highest 5% ($\geq$95th percentile) vs. average risk. Numbers in brackets are 95% confidence intervals. For subgroup analyses with less than 100 events of the endpoint, the box is marked '-'; these statistically less reliable results are reported in Table S2.

			Hazard ratios			
Group	**Clinical Endpoint**	**p**	**$HR_{80/20}$**	**$HR_{20/50}$**	**$HR_{80/50}$**	**$HR_{95/50}$**
All (n=590,750)	**Fatal Prostate Cancer**	$<10^{-16}$	4.42 [3.91-5.02]	0.48 [0.45-0.51]	2.12 [2.0-2.27]	3.0 [2.75-3.3]
Non-Hispanic White (n=420,473)		$<10^{-16}$	4.37 [3.77-5.05]	0.48 [0.45-0.52]	2.11 [1.96-2.27]	3.0 [2.69-3.34]
Black or African American (n=102,203)		$<10^{-16}$	2.37 [1.73-3.29]	0.66 [0.56-0.77]	1.57 [1.33-1.85]	1.9 [1.5-2.42]
All (n=590,750)	**Metastatic Prostate Cancer**	$<10^{-16}$	4.89 [4.57-5.21]	0.46 [0.44-0.47]	2.23 [2.16-2.31]	3.23 [3.07-3.39]
Non-Hispanic White (n=420,473)		$<10^{-16}$	4.62 [4.27-5.01]	0.47 [0.45-0.49]	2.17 [2.08-2.26]	3.13 [2.95-3.32]
Black or African American (n=102,203)		$<10^{-16}$	3.03 [2.62-3.51]	0.59 [0.55-0.63]	1.78 [1.65-1.92]	2.28 [2.05-2.54]
Hispanic White (n=27,651)		$<10^{-7}$	2.92 [1.96-4.43]	0.58 [0.47-0.71]	1.71 [1.4-2.11]	2.22 [1.65-3.01]
All (n=590,750)	**Prostate Cancer**	$<10^{-16}$	5.2 [5.09-5.31]	0.44 [0.44-0.45]	2.31 [2.28-2.33]	3.39 [3.33-3.44]
Non-Hispanic White (n=420,473)		$<10^{-16}$	5.12 [4.99-5.26]	0.45 [0.44-0.45]	2.29 [2.26-2.32]	3.38 [3.31-3.44]
Black or African American (n=102,203)		$<10^{-16}$	3.22 [3.06-3.39]	0.57 [0.56-0.58]	1.83 [1.79-1.88]	2.39 [2.3-2.48]
Hispanic White (n=27,651)		$<10^{-16}$	4.24 [3.74-4.85]	0.49 [0.45-0.52]	2.06 [1.93-2.2]	2.92 [2.66-3.22]
Asian (n=6,644)		$<10^{-16}$	5.32 [4.04-7.09]	0.43 [0.37-0.49]	2.27 [1.98-2.61]	3.49 [2.83-4.34]
Native American (n=5,835)		$<10^{-16}$	5.1 [4.04-6.55]	0.45 [0.4-0.51]	2.31 [2.05-2.62]	3.4 [2.85-4.12]
Pacific Islander (n=3,246)		$<10^{-10}$	3.58 [2.5-5.37]	0.53 [0.43-0.63]	1.9 [1.57-2.33]	2.6 [1.98-3.57]
Unknown (n=16,472)		$<10^{-16}$	5.27 [4.39-6.3]	0.44 [0.4-0.48]	2.32 [2.12-2.55]	3.45 [3.0-3.94]
Other (n=8,226)		$<10^{-16}$	5.52 [4.45-6.89]	0.43 [0.39-0.48]	2.36 [2.12-2.64]	3.59 [3.06-4.25]

Table 3.6 Multivariable Models Combining Self-Reported Race/Ethnicity, Family History (FH), and PHS290 for Three Prostate Cancer Clinical Endpoints Cox proportional hazards results for association with age at death from prostate cancer, at diagnosis of metastatic prostate cancer, and age at diagnosis with prostate cancer. *P*-values reported are from multivariable models using self-reported race/ethnicity, family history, and PHS290. For PHS290, effect size was illustrated via the hazard ratio ($HR_{80/20}$) for the highest 20% vs. lowest 20% of genetic risk. Hazard ratios for race/ethnicity were estimated using Non-Hispanic White as the reference. Hazard ratios for family history were for one or more first-degree relatives diagnosed with prostate cancer. This multivariable analysis was limited to the 375,763 participants who provided family history information in baseline survey data. Numbers in brackets are 95% confidence intervals. Significant predictors in the multivariable model are indicated by *($p<0.01$) and *** ($p<10^{-16}$).

	PHS290	Family History	Race/ Ethnicity						
Clinical Endpoints	$HR_{80/20}$	Family History	Black or African American	Hispanic White	Asian	Native American	Other	Pacific Islander	Unknown
Fatal Prostate Cancer	4.17 [3.59-4.88]***	1.67 [1.4-1.96]*	1.97 [1.69-2.31]***	1.04 [0.69-1.42]	0.68 [0.15-1.43]	0.98 [0.32-1.85]	1.9 [1.12-2.81]*	1.74 [0.0-4.43]	0.73 [0.0-1.95]
Metastatic Prostate Cancer	4.15 [3.81-4.53]***	1.53 [1.38-1.68]***	2.24 [2.07-2.42]***	1.32 [1.1-1.57]*	1.31 [0.81-1.85]	1.4 [0.97-1.86]	1.67 [1.25-2.08]*	1.14 [0.23-2.29]	1.52 [0.79-2.45]
Prostate Cancer	4.69 [4.57-4.81]***	1.78 [1.73-1.83]***	1.83 [1.78-1.87]***	1.1 [1.04-1.16]*	1.28 [1.13-1.43]*	1.04 [0.95-1.15]	1.2 [1.1-1.29]*	0.98 [0.73-1.25]	0.9 [0.72-1.08]

3.3.8 Author Contributions

Original Concept: Meghana Pagadala, Tyler Seibert

Project Supervision: Hannah Carter, Guneet Kaur, Brent Rose, Richard Hauger, Julie Lynch, Matt Panizzon,

Prostate Cancer Phenotyping: Julie Lynch, Patrick Alba, Kyung Lee, Fatai Agiri, Tori Anglin, Rishi Deka

Genetic Analyses: Meghana Pagadala, Tyler Seibert, Roshan Karunamuni

Manuscript Writing: Meghana Pagadala, Tyler Seibert

3.3.9 Acknowledgements

This research used data from the Million Veteran Program, Office of Research and Development, Veterans Health Administration. This research was supported by the Million Veteran Program MVP022 award # I01 CX001727 (PI: Richard L. Hauger MD).

This chapter in full is a reformatted preprint as it appears as "Polygenic risk of any, metastatic, and fatal prostate cancer in the Million Veteran Program" in *medRxiv* by Meghana Pagadala, Julie Lynch, Roshan Karunamuni, Patrick Alba, Kyung Lee, Fatai Agiri, Tori Anglin, Hannah Carter, Michael Gaziano, Guneet Jasuja, Rishi Deka, Brent Rose, Matthew Panizzon, Richard Hauger, Tyler Seibert.

3.3.10 References

1. A. Torkamani, N. E. Wineinger, E. J. Topol, The personal and clinical utility of polygenic risk scores. *Nat. Rev. Genet.* **19**, 581–590 (2018).

2. F. H. Schröder, J. Hugosson, M. J. Roobol, T. L. J. Tammela, M. Zappa, V. Nelen, M. Kwiatkowski, M. Lujan, L. Määttänen, H. Lilja, L. J. Denis, F. Recker, A. Paez, C. H. Bangma, S. Carlsson, D. Puliti, A. Villers, X. Rebillard, M. Hakama, U.-H. Stenman, P. Kujala, K. Taari, G. Aus, A. Huber, T. H. van der Kwast, R. H. N. van Schaik, H. J. de Koning, S. M. Moss, A. Auvinen, ERSPC Investigators, Screening and prostate cancer mortality: results of the European Randomised Study of Screening for Prostate Cancer (ERSPC) at 13 years of follow-up. *Lancet*. **384**, 2027–2035 (2014).

3. T. Callender, M. Emberton, S. Morris, R. Eeles, Z. Kote-Jarai, P. D. P. Pharoah, N. Pashayan, Polygenic risk-tailored screening for prostate cancer: A benefit-harm and cost-effectiveness modelling study. *PLoS Med.* **16**, e1002998 (2019).

4. S. Loeb, M. A. Bjurlin, J. Nicholson, T. L. Tammela, D. F. Penson, H. B. Carter, P. Carroll, R. Etzioni, Overdiagnosis and overtreatment of prostate cancer. *Eur. Urol.* **65**, 1046–1055 (2014).

5. L. A. Mucci, J. B. Hjelmborg, J. R. Harris, K. Czene, D. J. Havelick, T. Scheike, R. E. Graff, K. Holst, S. Möller, R. H. Unger, C. McIntosh, E. Nuttall, I. Brandt, K. L. Penney, M. Hartman, P. Kraft, G. Parmigiani, K. Christensen, M. Koskenvuo, N. V. Holm, K. Heikkilä, E. Pukkala, A. Skytthe, H.-O. Adami, J. Kaprio, Nordic Twin Study of Cancer (NorTwinCan) Collaboration, Familial risk and heritability of cancer among twins in Nordic countries. *JAMA*. **315**, 68–76 (2016).

6. T. Callender, M. Emberton, S. Morris, P. D. P. Pharoah, N. Pashayan, Benefit, harm, and cost-effectiveness associated with magnetic resonance imaging before biopsy in age-based and risk-stratified screening for prostate cancer. *JAMA Netw. Open*. **4**, e2037657 (2021).

7. T. M. Seibert, C. C. Fan, Y. Wang, V. Zuber, R. Karunamuni, J. K. Parsons, R. A. Eeles, D. F. Easton, Z. Kote-Jarai, A. A. Al Olama, S. B. Garcia, K. Muir, H. Grönberg, F. Wiklund, M. Aly, J. Schleutker, C. Sipeky, T. L. Tammela, B. G. Nordestgaard, S. F. Nielsen, M. Weischer, R. Bisbjerg, M. A. Røder, P. Iversen, T. J. Key, R. C. Travis, D. E. Neal, J. L. Donovan, F. C. Hamdy, P. Pharoah, N. Pashayan, K.-T. Khaw, C. Maier, W. Vogel, M. Luedeke, K. Herkommer, A. S. Kibel, C. Cybulski, D. Wokolorczyk, W. Kluzniak, L. Cannon-Albright, H. Brenner, K. Cuk, K.-U. Saum, J. Y. Park, T. A. Sellers, C. Slavov, R. Kaneva, V. Mitev, J. Batra, J. A. Clements, A. Spurdle, M. R. Teixeira, P. Paulo, S. Maia, H. Pandha, A. Michael, A. Kierzek, D. S. Karow, I. G. Mills, O. A. Andreassen, A. M. Dale, PRACTICAL Consortium*, Polygenic hazard score to guide screening for aggressive prostate cancer: development and validation in large scale cohorts. *BMJ*. **360**, j5757 (2018).

8. F. R. Schumacher, A. A. A. Olama, S. I. Berndt, S. Benlloch, M. Ahmed, E. J. Saunders, T. Dadaev, D. Leongamornlert, E. Anokian, C. Cieza-Borrella, C. Goh, M. N. Brook, X. Sheng, L. Fachal, J. Dennis, J. Tyrer, K. Muir, A. Lophatananon, V. L. Stevens, S. M. Gapstur, B.

D. Carter, C. M. Tangen, P. J. Goodman, I. M. Thompson Jr, J. Batra, S. Chambers, L. Moya, J. Clements, L. Horvath, W. Tilley, G. P. Risbridger, H. Gronberg, M. Aly, T. Nordström, P. Pharoah, N. Pashayan, J. Schleutker, T. L. J. Tammela, C. Sipeky, A. Auvinen, D. Albanes, S. Weinstein, A. Wolk, N. Håkansson, C. M. L. West, A. M. Dunning, N. Burnet, L. A. Mucci, E. Giovannucci, G. L. Andriole, O. Cussenot, G. Cancel-Tassin, S. Koutros, L. E. Beane Freeman, K. D. Sorensen, T. F. Orntoft, M. Borre, L. Maehle, E. M. Grindedal, D. E. Neal, J. L. Donovan, F. C. Hamdy, R. M. Martin, R. C. Travis, T. J. Key, R. J. Hamilton, N. E. Fleshner, A. Finelli, S. A. Ingles, M. C. Stern, B. S. Rosenstein, S. L. Kerns, H. Ostrer, Y.-J. Lu, H.-W. Zhang, N. Feng, X. Mao, X. Guo, G. Wang, Z. Sun, G. G. Giles, M. C. Southey, R. J. MacInnis, L. M. FitzGerald, A. S. Kibel, B. F. Drake, A. Vega, A. Gómez-Caamaño, R. Szulkin, M. Eklund, M. Kogevinas, J. Llorca, G. Castaño-Vinyals, K. L. Penney, M. Stampfer, J. Y. Park, T. A. Sellers, H.-Y. Lin, J. L. Stanford, C. Cybulski, D. Wokolorczyk, J. Lubinski, E. A. Ostrander, M. S. Geybels, B. G. Nordestgaard, S. F. Nielsen, M. Weischer, R. Bisbjerg, M. A. Røder, P. Iversen, H. Brenner, K. Cuk, B. Holleczek, C. Maier, M. Luedeke, T. Schnoeller, J. Kim, C. J. Logothetis, E. M. John, M. R. Teixeira, P. Paulo, M. Cardoso, S. L. Neuhausen, L. Steele, Y. C. Ding, K. De Ruyck, G. De Meerleer, P. Ost, A. Razack, J. Lim, S.-H. Teo, D. W. Lin, L. F. Newcomb, D. Lessel, M. Gamulin, T. Kulis, R. Kaneva, N. Usmani, S. Singhal, C. Slavov, V. Mitev, M. Parliament, F. Claessens, S. Joniau, T. Van den Broeck, S. Larkin, P. A. Townsend, C. Aukim-Hastie, M. Gago-Dominguez, J. E. Castelao, M. E. Martinez, M. J. Roobol, G. Jenster, R. H. N. van Schaik, F. Menegaux, T. Truong, Y. A. Koudou, J. Xu, K.-T. Khaw, L. Cannon-Albright, H. Pandha, A. Michael, S. N. Thibodeau, S. K. McDonnell, D. J. Schaid, S. Lindstrom, C. Turman, J. Ma, D. J. Hunter, E. Riboli, A. Siddiq, F. Canzian, L. N. Kolonel, L. Le Marchand, R. N. Hoover, M. J. Machiela, Z. Cui, P. Kraft, C. I. Amos, D. V. Conti, D. F. Easton, F. Wiklund, S. J. Chanock, B. E. Henderson, Z. Kote-Jarai, C. A. Haiman, R. A. Eeles, Profile Study, Australian Prostate Cancer BioResource (APCB), IMPACT Study, Canary PASS Investigators, Breast and Prostate Cancer Cohort Consortium (BPC3), PRACTICAL (Prostate Cancer Association Group to Investigate Cancer-Associated Alterations in the Genome) Consortium, Cancer of the Prostate in Sweden (CAPS), Prostate Cancer Genome-wide Association Study of Uncommon Susceptibility Loci (PEGASUS), Genetic Associations and Mechanisms in Oncology (GAME-ON)/Elucidating Loci Involved in Prostate Cancer Susceptibility (ELLIPSE) Consortium, Author Correction: Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. *Nat. Genet.* **51**, 363 (2019).

9. M.-P. Huynh-Le, R. Karunamuni, C. C. Fan, W. K. Thompson, K. Muir, A. Lophatananon, K. Tye, A. Wolk, N. Håkansson, I. G. Mills, O. A. Andreassen, A. M. Dale, T. M. Seibert, PRACTICAL Consortium, Common genetic and clinical risk factors: association with fatal prostate cancer in the Cohort of Swedish Men. *Prostate Cancer Prostatic Dis.* (2021), doi:10.1038/s41391-021-00341-4.

10. R. A. Karunamuni, M.-P. Huynh-Le, C. C. Fan, W. Thompson, R. A. Eeles, Z. Kote-Jarai, K. Muir, A. Lophatananon, UKGPCS collaborators, J. Schleutker, N. Pashayan, J. Batra, APCB BioResource (Australian Prostate Cancer BioResource), H. Grönberg, E. I. Walsh, E. L. Turner, A. Lane, R. M. Martin, D. E. Neal, J. L. Donovan, F. C. Hamdy, B. G. Nordestgaard, C. M. Tangen, R. J. MacInnis, A. Wolk, D. Albanes, C. A. Haiman, R. C. Travis, J. L. Stanford, L. A. Mucci, C. M. L. West, S. F. Nielsen, A. S. Kibel, F. Wiklund, O. Cussenot, S. I. Berndt, S. Koutros, K. D. Sørensen, C. Cybulski, E. M. Grindedal, J. Y. Park, S. A.

Ingles, C. Maier, R. J. Hamilton, B. S. Rosenstein, A. Vega, IMPACT Study Steering Committee and Collaborators, M. Kogevinas, K. L. Penney, M. R. Teixeira, H. Brenner, E. M. John, R. Kaneva, C. J. Logothetis, S. L. Neuhausen, A. Razack, L. F. Newcomb, Canary PASS Investigators, M. Gamulin, N. Usmani, F. Claessens, M. Gago-Dominguez, P. A. Townsend, M. J. Roobol, W. Zheng, Profile Study Steering Committee, I. G. Mills, O. A. Andreassen, A. M. Dale, T. M. Seibert, PRACTICAL Consortium, Additional SNPs improve risk stratification of a polygenic hazard score for prostate cancer. *Prostate Cancer Prostatic Dis.* **24**, 532–541 (2021).

11. M.-P. Huynh-Le, T. Å. Myklebust, C. H. Feng, R. Karunamuni, T. B. Johannesen, A. M. Dale, O. A. Andreassen, T. M. Seibert, Age dependence of modern clinical risk groups for localized prostate cancer-A population-based study. *Cancer*. **126**, 1691–1699 (2020).

12. M.-P. Huynh-Le, C. C. Fan, R. Karunamuni, E. I. Walsh, E. L. Turner, J. A. Lane, R. M. Martin, D. E. Neal, J. L. Donovan, F. C. Hamdy, J. K. Parsons, R. A. Eeles, D. F. Easton, Z. Kote-Jarai, A. Amin Al Olama, S. Benlloch Garcia, K. Muir, H. Grönberg, F. Wiklund, M. Aly, J. Schleutker, C. Sipeky, T. L. Tammela, B. G. Nordestgaard, T. J. Key, R. C. Travis, P. D. P. Pharoah, N. Pashayan, K.-T. Khaw, S. N. Thibodeau, S. K. McDonnell, D. J. Schaid, C. Maier, W. Vogel, M. Luedeke, K. Herkommer, A. S. Kibel, C. Cybulski, D. Wokolorczyk, W. Kluzniak, L. A. Cannon-Albright, H. Brenner, B. Schöttker, B. Holleczek, J. Y. Park, T. A. Sellers, H.-Y. Lin, C. K. Slavov, R. P. Kaneva, V. I. Mitev, J. Batra, J. A. Clements, A. B. Spurdle, M. R. Teixeira, P. Paulo, S. Maia, H. Pandha, A. Michael, I. G. Mills, O. A. Andreassen, A. M. Dale, T. M. Seibert, Australian Prostate Cancer BioResource (APCB), PRACTICAL Consortium, A genetic risk score to personalize Prostate Cancer screening, applied to population data. *Cancer Epidemiol. Biomarkers Prev.* **29**, 1731–1738 (2020).

13. L. Duncan, H. Shen, B. Gelaye, J. Meijsen, K. Ressler, M. Feldman, R. Peterson, B. Domingue, Analysis of polygenic risk score usage and performance in diverse human populations. *Nat. Commun.* **10**, 3328 (2019).

14. S. Petrovski, D. B. Goldstein, Unequal representation of genetic variation across ancestry groups creates healthcare inequality in the application of precision medicine. *Genome Biol.* **17** (2016), doi:10.1186/s13059-016-1016-y.

15. A. R. Martin, M. Kanai, Y. Kamatani, Y. Okada, B. M. Neale, M. J. Daly, Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).

16. A. B. Popejoy, S. M. Fullerton, Genomics is failing on diversity. *Nature*. **538**, 161–164 (2016).

17. C. E. DeSantis, K. D. Miller, A. Goding Sauer, A. Jemal, R. L. Siegel, Cancer statistics for African Americans, 2019. *CA Cancer J. Clin.* **69**, 211–233 (2019).

18. A. Tsodikov, R. Gulati, T. M. de Carvalho, E. A. M. Heijnsdijk, R. A. Hunter-Merrill, A. B. Mariotto, H. J. de Koning, R. Etzioni, Is prostate cancer different in black men? Answers from 3 natural history models. *Cancer*. **123**, 2312–2319 (2017).

19. R. A. Karunamuni, M.-P. Huynh-Le, C. C. Fan, W. Thompson, R. A. Eeles, Z. Kote-Jarai, K. Muir, UKGPCS Collaborators, A. Lophatananon, C. M. Tangen, P. J. Goodman, I. M. Thompson Jr, W. J. Blot, W. Zheng, A. S. Kibel, B. F. Drake, O. Cussenot, G. Cancel-Tassin, F. Menegaux, T. Truong, J. Y. Park, H.-Y. Lin, J. T. Bensen, E. T. H. Fontham, J. L. Mohler, J. A. Taylor, L. Multigner, P. Blanchet, L. Brureau, M. Romana, R. J. Leach, E. M. John, J. Fowke, W. S. Bush, M. Aldrich, D. C. Crawford, S. Srivastava, J. C. Cullen, G. Petrovics, M.-É. Parent, J. J. Hu, M. Sanderson, I. G. Mills, O. A. Andreassen, A. M. Dale, T. M. Seibert, PRACTICAL Consortium, African-specific improvement of a polygenic hazard score for age at diagnosis of prostate cancer. *Int. J. Cancer*. **148**, 99–105 (2021).

20. R. A. Karunamuni, M.-P. Huynh-Le, C. C. Fan, W. Thompson, A. Lui, M. E. Martinez, B. S. Rose, B. Mahal, R. A. Eeles, Z. Kote-Jarai, K. Muir, A. Lophatananon, UKGPCS Collaborators, C. M. Tangen, P. J. Goodman, I. M. Thompson Jr, W. J. Blot, W. Zheng, A. S. Kibel, B. F. Drake, O. Cussenot, G. Cancel-Tassin, F. Menegaux, T. Truong, J. Y. Park, H.-Y. Lin, J. A. Taylor, J. T. Bensen, J. L. Mohler, E. T. H. Fontham, L. Multigner, P. Blanchet, L. Brureau, M. Romana, R. J. Leach, E. M. John, J. H. Fowke, W. S. Bush, M. C. Aldrich, D. C. Crawford, J. Cullen, G. Petrovics, M.-É. Parent, J. J. Hu, M. Sanderson, PRACTICAL Consortium, I. G. Mills, O. A. Andreassen, A. M. Dale, T. M. Seibert, Performance of African-ancestry-specific polygenic hazard score varies according to local ancestry in 8q24. *Prostate Cancer Prostatic Dis.* (2021), doi:10.1038/s41391-021-00403-7.

21. D. V. Conti, B. F. Darst, L. C. Moss, E. J. Saunders, X. Sheng, A. Chou, F. R. Schumacher, A. A. A. Olama, S. Benlloch, T. Dadaev, M. N. Brook, A. Sahimi, T. J. Hoffmann, A. Takahashi, K. Matsuda, Y. Momozawa, M. Fujita, K. Muir, A. Lophatananon, P. Wan, L. Le Marchand, L. R. Wilkens, V. L. Stevens, S. M. Gapstur, B. D. Carter, J. Schleutker, T. L. J. Tammela, C. Sipeky, A. Auvinen, G. G. Giles, M. C. Southey, R. J. MacInnis, C. Cybulski, D. Wokołorczyk, J. Lubiński, D. E. Neal, J. L. Donovan, F. C. Hamdy, R. M. Martin, B. G. Nordestgaard, S. F. Nielsen, M. Weischer, S. E. Bojesen, M. A. Røder, P. Iversen, J. Batra, S. Chambers, L. Moya, L. Horvath, J. A. Clements, W. Tilley, G. P. Risbridger, H. Gronberg, M. Aly, R. Szulkin, M. Eklund, T. Nordström, N. Pashayan, A. M. Dunning, M. Ghoussaini, R. C. Travis, T. J. Key, E. Riboli, J. Y. Park, T. A. Sellers, H.-Y. Lin, D. Albanes, S. J. Weinstein, L. A. Mucci, E. Giovannucci, S. Lindstrom, P. Kraft, D. J. Hunter, K. L. Penney, C. Turman, C. M. Tangen, P. J. Goodman, I. M. Thompson Jr, R. J. Hamilton, N. E. Fleshner, A. Finelli, M.-É. Parent, J. L. Stanford, E. A. Ostrander, M. S. Geybels, S. Koutros, L. E. B. Freeman, M. Stampfer, A. Wolk, N. Håkansson, G. L. Andriole, R. N. Hoover, M. J. Machiela, K. D. Sørensen, M. Borre, W. J. Blot, W. Zheng, E. D. Yeboah, J. E. Mensah, Y.-J. Lu, H.-W. Zhang, N. Feng, X. Mao, Y. Wu, S.-C. Zhao, Z. Sun, S. N. Thibodeau, S. K. McDonnell, D. J. Schaid, C. M. L. West, N. Burnet, G. Barnett, C. Maier, T. Schnoeller, M. Luedeke, A. S. Kibel, B. F. Drake, O. Cussenot, G. Cancel-Tassin, F. Menegaux, T. Truong, Y. A. Koudou, E. M. John, E. M. Grindedal, L. Maehle, K.-T. Khaw, S. A. Ingles, M. C. Stern, A. Vega, A. Gómez-Caamaño, L. Fachal, B. S. Rosenstein, S. L. Kerns, H. Ostrer, M. R. Teixeira, P. Paulo, A. Brandão, S. Watya, A. Lubwama, J. T. Bensen, E. T. H. Fontham, J. Mohler, J. A. Taylor, M. Kogevinas, J. Llorca, G. Castaño-Vinyals, L. Cannon-Albright, C. C. Teerlink, C. D. Huff, S. S. Strom, L. Multigner, P. Blanchet, L. Brureau, R. Kaneva, C. Slavov, V. Mitev, R. J. Leach, B. Weaver, H. Brenner, K. Cuk, B. Holleczek, K.-U. Saum, E. A. Klein, A. W. Hsing, R. A. Kittles, A. B. Murphy, C. J. Logothetis, J. Kim, S. L. Neuhausen, L. Steele, Y. C. Ding, W. B. Isaacs, B. Nemesure, A. J. M. Hennis, J. Carpten, H. Pandha, A.

Michael, K. De Ruyck, G. De Meerleer, P. Ost, J. Xu, A. Razack, J. Lim, S.-H. Teo, L. F. Newcomb, D. W. Lin, J. H. Fowke, C. Neslund-Dudas, B. A. Rybicki, M. Gamulin, D. Lessel, T. Kulis, N. Usmani, S. Singhal, M. Parliament, F. Claessens, S. Joniau, T. Van den Broeck, M. Gago-Dominguez, J. E. Castelao, M. E. Martinez, S. Larkin, P. A. Townsend, C. Aukim-Hastie, W. S. Bush, M. C. Aldrich, D. C. Crawford, S. Srivastava, J. C. Cullen, G. Petrovics, G. Casey, M. J. Roobol, G. Jenster, R. H. N. van Schaik, J. J. Hu, M. Sanderson, R. Varma, R. McKean-Cowdin, M. Torres, N. Mancuso, S. I. Berndt, S. K. Van Den Eeden, D. F. Easton, S. J. Chanock, M. B. Cook, F. Wiklund, H. Nakagawa, J. S. Witte, R. A. Eeles, Z. Kote-Jarai, C. A. Haiman, 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).

22. M.-P. Huynh-Le, C. C. Fan, R. Karunamuni, W. K. Thompson, M. E. Martinez, R. A. Eeles, Z. Kote-Jarai, K. Muir, J. Schleutker, N. Pashayan, J. Batra, H. Grönberg, D. E. Neal, J. L. Donovan, F. C. Hamdy, R. M. Martin, S. F. Nielsen, B. G. Nordestgaard, F. Wiklund, C. M. Tangen, G. G. Giles, A. Wolk, D. Albanes, R. C. Travis, W. J. Blot, W. Zheng, M. Sanderson, J. L. Stanford, L. A. Mucci, C. M. L. West, A. S. Kibel, O. Cussenot, S. I. Berndt, S. Koutros, K. D. Sørensen, C. Cybulski, E. M. Grindedal, F. Menegaux, K.-T. Khaw, J. Y. Park, S. A. Ingles, C. Maier, R. J. Hamilton, S. N. Thibodeau, B. S. Rosenstein, Y.-J. Lu, S. Watya, A. Vega, M. Kogevinas, K. L. Penney, C. Huff, M. R. Teixeira, L. Multigner, R. J. Leach, L. Cannon-Albright, H. Brenner, E. M. John, R. Kaneva, C. J. Logothetis, S. L. Neuhausen, K. De Ruyck, H. Pandha, A. Razack, L. F. Newcomb, J. H. Fowke, M. Gamulin, N. Usmani, F. Claessens, M. Gago-Dominguez, P. A. Townsend, W. S. Bush, M. J. Roobol, M.-É. Parent, J. J. Hu, I. G. Mills, O. A. Andreassen, A. M. Dale, T. M. Seibert, UKGPCS collaborators, APCB (Australian Prostate Cancer BioResource), NC-LA PCaP Investigators, IMPACT Study Steering Committee and Collaborators, Canary PASS Investigators, Profile Study Steering Committee, PRACTICAL Consortium, Polygenic hazard score is associated with prostate cancer in multi-ethnic populations. *Nat. Commun.* **12**, 1236 (2021).

23. M.-P. Huynh-Le, R. Karunamuni, C. C. Fan, L. Asona, W. K. Thompson, M. E. Martinez, R. A. Eeles, Z. Kote-Jarai, K. R. Muir, A. Lophatananon, J. Schleutker, N. Pashayan, J. Batra, H. Grönberg, D. E. Neal, B. G. Nordestgaard, C. M. Tangen, R. J. MacInnis, A. Wolk, D. Albanes, C. A. Haiman, R. C. Travis, W. J. Blot, J. L. Stanford, L. A. Mucci, C. M. L. West, S. F. Nielsen, A. S. Kibel, O. Cussenot, S. I. Berndt, S. Koutros, K. D. Sørensen, C. Cybulski, E. M. Grindedal, F. Menegaux, J. Y. Park, S. A. Ingles, C. Maier, R. J. Hamilton, B. S. Rosenstein, Y.-J. Lu, S. Watya, A. Vega, M. Kogevinas, F. Wiklund, K. L. Penney, C. D. Huff, M. R. Teixeira, L. Multigner, R. J. Leach, H. Brenner, E. M. John, R. Kaneva, C. J. Logothetis, S. L. Neuhausen, K. De Ruyck, P. Ost, A. Razack, L. F. Newcomb, J. H. Fowke, M. Gamulin, A. Abraham, F. Claessens, J. E. Castelao, P. A. Townsend, D. C. Crawford, G. Petrovics, R. H. N. van Schaik, M.-É. Parent, J. J. Hu, W. Zheng, UKGPCS collaborators, APCB (Australian Prostate Cancer BioResource), NC-LA PCaP Investigators, IMPACT Study Steering Committee and Collaborators, Canary PASS Investigators, Profile Study Steering Committee, PRACTICAL Consortium, I. G. Mills, O. A. Andreassen, A. M. Dale, T. M. Seibert, Prostate cancer risk stratification improvement across multiple ancestries with new polygenic hazard score. *Prostate Cancer Prostatic Dis.* (2022), doi:10.1038/s41391-022-00497-7.

24. M.-P. Huynh-Le, R. Karunamuni, C. C. Fan, A. Lui, W. K. Thompson, M. E. Martinez, R. A. Eeles, Z. Kote-Jarai, K. R. Muir, A. Lophatananon, J. Schleutker, N. Pashayan, J. Batra, H. Gronberg, D. E. Neal, B. G. Nordestgaard, C. M. Tangen, R. J. MacInnis, A. Wolk, D. Albanes, C. A. Haiman, R. C. Travis, W. J. Blot, J. L. Stanford, L. A. Mucci, C. M. L. West, S. F. Nielsen, A. S. Kibel, O. Cussenot, S. I. Berndt, S. Koutros, K. D. Soerensen, C. Cybulski, E. M. Grindedal, F. Menegaux, J. Y. Park, S. A. Ingles, C. Maier, R. J. Hamilton, B. S. Rosenstien, Y.-J. Lu, S. Watya, A. Vega, M. Kogevinas, F. Wiklund, K. L. Penney, C. D. Huff, M. R. Teixeira, L. Multigner, R. J. Leach, H. Brenner, E. M. John, R. Kaneva, C. J. Logothetis, S. L. Neuhausen, K. De Ruyck, P. Ost, A. Razack, L. F. Newcomb, J. H. Fowke, M. Gamulin, A. Abraham, F. Claessens, J. E. Castelao, P. A. Townsend, D. C. Crawford, G. Petrovics, R. H. N. van Schaik, M.-E. Parent, J. J. Hu, W. Zheng, I. G. Mills, O. A. Andreassen, A. M. Dale, T. M. Seibert, UKGPCS collaborators, Australian Prostate Cancer BioResource, NC-LA PCaP Investigators, The IMPACT Study Steering Committee and Collaborators, Canary PASS Investigators, The Profile Study Steering Committee, The PRACTICAL Consortium, Prostate cancer risk stratification improved across multiple ancestries with new polygenic hazard score (2021), , doi:10.1101/2021.08.14.21261931.

25. J. M. Gaziano, J. Concato, M. Brophy, L. Fiore, S. Pyarajan, J. Breeling, S. Whitbourne, J. Deen, C. Shannon, D. Humphries, P. Guarino, M. Aslan, D. Anderson, R. LaFleur, T. Hammond, K. Schaa, J. Moser, G. Huang, S. Muralidhar, R. Przygodzki, T. J. O'Leary, Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *J. Clin. Epidemiol.* **70**, 214–223 (2016).

26. H. Hunter-Zinck, Y. Shi, M. Li, B. R. Gorman, S.-G. Ji, N. Sun, T. Webster, A. Liem, P. Hsieh, P. Devineni, P. Karnam, X. Gong, L. Radhakrishnan, J. Schmidt, T. L. Assimes, J. Huang, C. Pan, D. Humphries, M. Brophy, J. Moser, S. Muralidhar, G. D. Huang, R. Przygodzki, J. Concato, J. M. Gaziano, J. Gelernter, C. J. O'Donnell, E. R. Hauser, H. Zhao, T. J. O'Leary, VA Million Veteran Program, P. S. Tsao, S. Pyarajan, Genotyping array design and data quality control in the Million Veteran Program. *Am. J. Hum. Genet.* **106**, 535–548 (2020).

27. P. R. Alba, A. Gao, K. M. Lee, T. Anglin-Foote, B. Robison, E. Katsoulakis, B. S. Rose, O. Efimova, J. P. Ferraro, O. V. Patterson, J. B. Shelton, S. L. Duvall, J. A. Lynch, Ascertainment of Veterans with Metastatic Prostate Cancer in Electronic Health Records: Demonstrating the Case for Natural Language Processing. *JCO Clin Cancer Inform*. **5**, 1005–1014 (2021).

28. R. A. Karunamuni, M.-P. Huynh-Le, C. C. Fan, R. A. Eeles, D. F. Easton, Z. Kote-Jarai, A. Amin Al Olama, S. Benlloch Garcia, K. Muir, H. Gronberg, F. Wiklund, M. Aly, J. Schleutker, C. Sipeky, T. L. J. Tammela, B. G. Nordestgaard, T. J. Key, R. C. Travis, D. E. Neal, J. L. Donovan, F. C. Hamdy, P. Pharoah, N. Pashayan, K.-T. Khaw, S. N. Thibodeau, S. K. McDonnell, D. J. Schaid, C. Maier, W. Vogel, M. Luedeke, K. Herkommer, A. S. Kibel, C. Cybulski, D. Wokolorczyk, W. Kluzniak, L. Cannon-Albright, H. Brenner, B. Schöttker, B. Holleczek, J. Y. Park, T. A. Sellers, H.-Y. Lin, C. Slavov, R. Kaneva, V. Mitev, J. Batra, J. A. Clements, A. Spurdle, Australian Prostate Cancer BioResource (APCB), M. R. Teixeira, P. Paulo, S. Maia, H. Pandha, A. Michael, I. G. Mills, O. A. Andreassen, A. M. Dale, T. M. Seibert, PRACTICAL Consortium, The effect of sample size on polygenic hazard models for prostate cancer. *Eur. J. Hum. Genet.* **28**, 1467–1475 (2020).

29. F. R. Wendt, G. A. Pathak, J. Vahey, X. Qin, D. Koller, B. Cabrera-Mendoza, A. Haeny, K. M. Harrington, N. Rajeevan, L. M. Duong, D. F. Levey, F. De Angelis, A. De Lillo, T. B. Bigdeli, S. Pyarajan, J. M. Gaziano, J. Gelernter, M. Aslan, D. Provenzale, D. A. Helmer, E. R. Hauser, R. Polimanti, VA Million Veteran Program, Department of Veteran Affairs Cooperative Study Program (#2006), Modeling the longitudinal changes of ancestry diversity in the Million Veteran Program. *bioRxiv* (2022), , doi:10.1101/2022.01.24.477583.

30. The 1000 Genomes Project Consortium, A. Auton, G. R. Abecasis, D. M. Altshuler (Co-Chair), R. M. Durbin (Co-Chair), G. R. Abecasis, D. R. Bentley, A. Chakravarti, A. G. Clark, P. Donnelly, E. E. Eichler, P. Flicek, S. B. Gabriel, R. A. Gibbs, E. D. Green, M. E. Hurles, B. M. Knoppers, J. O. Korbel, E. S. Lander, C. Lee, H. Lehrach, E. R. Mardis, G. T. Marth, G. A. McVean, D. A. Nickerson, J. P. Schmidt, S. T. Sherry, J. Wang, R. K. Wilson, R. A. Gibbs (Principal, E. Boerwinkle, H. Doddapaneni, Y. Han, V. Korchina, C. Kovar, S. Lee, D. Muzny, J. G. Reid, Y. Zhu, J. Wang (Principal, Y. Chang, Q. Feng, X. Fang, X. Guo, M. Jian, H. Jiang, X. Jin, T. Lan, G. Li, J. Li, Y. Li, S. Liu, X. Liu, Y. Lu, X. Ma, M. Tang, B. Wang, G. Wang, H. Wu, R. Wu, X. Xu, Y. Yin, D. Zhang, W. Zhang, J. Zhao, M. Zhao, X. Zheng, E. S. Lander (Principal, D. M. Altshuler, S. B. Gabriel (Co-Chair), N. Gupta, N. Gharani, L. H. Toji, N. P. Gerry, A. M. Resch, P. Flicek (Principal, J. Barker, L. Clarke, L. Gil, S. E. Hunt, G. Kelman, E. Kulesha, R. Leinonen, W. M. McLaren, R. Radhakrishnan, A. Roa, D. Smirnov, R. E. Smith, I. Streeter, A. Thormann, I. Toneva, B. Vaughan, X. Zheng-Bradley, D. R. Bentley (Principal, R. Grocock, S. Humphray, T. James, Z. Kingsbury, H. Lehrach (Principal, R. Sudbrak (Project L, M. W. Albrecht, V. S. Amstislavskiy, T. A. Borodina, M. Lienhard, F. Mertes, M. Sultan, B. Timmermann, M.-L. Yaspo, E. R. Mardis (Co-Princi, R. K. Wilson (Co-Princi, L. Fulton, R. Fulton, S. T. Sherry (Principal, V. Ananiev, Z. Belaia, D. Beloslyudtsev, N. Bouk, C. Chen, D. Church, R. Cohen, C. Cook, J. Garner, T. Hefferon, M. Kimelman, C. Liu, J. Lopez, P. Meric, C. O'Sullivan, Y. Ostapchuk, L. Phan, S. Ponomarov, V. Schneider, E. Shekhtman, K. Sirotkin, D. Slotta, H. Zhang, G. A. McVean (Principal, R. M. Durbin (Principal, S. Balasubramaniam, J. Burton, P. Danecek, T. M. Keane, A. Kolb-Kokocinski, S. McCarthy, J. Stalker, M. Quail, J. P. Schmidt (Principal, C. J. Davies, J. Gollub, T. Webster, B. Wong, Y. Zhan, A. Auton (Principal, C. L. Campbell, Y. Kong, A. Marcketta, R. A. Gibbs (Principal, F. Yu (Project L, L. Antunes, M. Bainbridge, D. Muzny, A. Sabo, Z. Huang, J. Wang (Principal, L. J. M. Coin, L. Fang, X. Guo, X. Jin, G. Li, Q. Li, Y. Li, Z. Li, H. Lin, B. Liu, R. Luo, H. Shao, Y. Xie, C. Ye, C. Yu, F. Zhang, H. Zheng, H. Zhu, C. Alkan, E. Dal, F. Kahveci, G. T. Marth (Principal, E. P. Garrison (Project L, D. Kural, W.-P. Lee, W. Fung Leong, M. Stromberg, A. N. Ward, J. Wu, M. Zhang, M. J. Daly (Principal, M. A. DePristo (Project L, R. E. Handsaker (Project L, D. M. Altshuler, E. Banks, G. Bhatia, G. del Angel, S. B. Gabriel, G. Genovese, N. Gupta, H. Li, S. Kashin, E. S. Lander, S. A. McCarroll, J. C. Nemesh, R. E. Poplin, S. C. Yoon (Principal, J. Lihm, V. Makarov, A. G. Clark (Principal, S. Gottipati, A. Keinan, J. L. Rodriguez-Flores, J. O. Korbel (Principal, T. Rausch (Project L, M. H. Fritz, A. M. Stütz, P. Flicek (Principal, K. Beal, L. Clarke, A. Datta, J. Herrero, W. M. McLaren, G. R. S. Ritchie, R. E. Smith, D. Zerbino, X. Zheng-Bradley, P. C. Sabeti (Principal, I. Shlyakhter, S. F. Schaffner, J. Vitti, D. N. Cooper (Principal, E. V. Ball, P. D. Stenson, D. R. Bentley (Principal, B. Barnes, M. Bauer, R. Keira Cheetham, A. Cox, M. Eberle, S. Humphray, S. Kahn, L. Murray, J. Peden, R. Shaw, E. E. Kenny (Principal, M. A. Batzer (Principal, M. K. Konkel, J. A. Walker, D. G. MacArthur (Principal, M. Lek, R. Sudbrak (Project L, V. S. Amstislavskiy, R. Herwig, E. R. Mardis (Co-Princi, L. Ding, D. C. Koboldt, D. Larson, K. Ye, S. Gravel, A. Swaroop, E. Chew, T.

Lappalainen (Principal, Y. Erlich (Principal, M. Gymrek, T. Frederick Willems, J. T. Simpson, M. D. Shriver (Principal, J. A. Rosenfeld (Principal, C. D. Bustamante (Principal, S. B. Montgomery (Principal, F. M. De La Vega (Principal, J. K. Byrnes, A. W. Carroll, M. K. DeGorter, P. Lacroute, B. K. Maples, A. R. Martin, A. Moreno-Estrada, S. S. Shringarpure, F. Zakharia, E. Halperin (Principal, Y. Baran, C. Lee (Principal, E. Cerveira, J. Hwang, A. Malhotra (Co-Projec, D. Plewczynski, K. Radew, M. Romanovitch, C. Zhang (Co-Projec, F. C. L. Hyland, D. W. Craig (Principal, A. Christoforides, N. Homer, T. Izatt, A. A. Kurdoglu, S. A. Sinari, K. Squire, S. T. Sherry (Principal, C. Xiao, J. Sebat (Principal, D. Antaki, M. Gujral, A. Noor, K. Ye, E. G. Burchard (Principal, R. D. Hernandez (Principal, C. R. Gignoux, D. Haussler (Principal, S. J. Katzman, W. James Kent, B. Howie, A. Ruiz-Linares (Principal, E. T. Dermitzakis (Principal, S. E. Devine (Principal, G. R. Abecasis (Principal, H. Min Kang (Project L, J. M. Kidd (Principal, T. Blackwell, S. Caron, W. Chen, S. Emery, L. Fritsche, C. Fuchsberger, G. Jun, B. Li, R. Lyons, C. Scheller, C. Sidore, S. Song, E. Sliwerska, D. Taliun, A. Tan, R. Welch, M. Kate Wing, X. Zhan, P. Awadalla (Principal, A. Hodgkinson, Y. Li, X. Shi (Principal, A. Quitadamo, G. Lunter (Principal, G. A. McVean (Principal, J. L. Marchini (Principal, S. Myers (Principal, C. Churchhouse, O. Delaneau, A. Gupta-Hinch, W. Kretzschmar, Z. Iqbal, I. Mathieson, A. Menelaou, A. Rimmer, D. K. Xifara, T. K. Oleksyk (Principal, Y. Fu (Principal, X. Liu, M. Xiong, L. Jorde (Principal, D. Witherspoon, J. Xing, E. E. Eichler (Principal, B. L. Browning (Principal, S. R. Browning (Principal, F. Hormozdiari, P. H. Sudmant, E. Khurana (Principal, R. M. Durbin (Principal, M. E. Hurles (Principal, C. Tyler-Smith (Principal, C. A. Albers, Q. Ayub, S. Balasubramaniam, Y. Chen, V. Colonna, P. Danecek, L. Jostins, T. M. Keane, S. McCarthy, K. Walter, Y. Xue, M. B. Gerstein (Principal, A. Abyzov, S. Balasubramanian, J. Chen, D. Clarke, Y. Fu, A. O. Harmanci, M. Jin, D. Lee, J. Liu, X. Jasmine Mu, J. Zhang, Y. Zhang, Y. Li, R. Luo, H. Zhu, C. Alkan, E. Dal, F. Kahveci, G. T. Marth (Principal, E. P. Garrison, D. Kural, W.-P. Lee, A. N. Ward, J. Wu, M. Zhang, S. A. McCarroll (Principal, R. E. Handsaker (Project L, D. M. Altshuler, E. Banks, G. del Angel, G. Genovese, C. Hartl, H. Li, S. Kashin, J. C. Nemesh, K. Shakir, S. C. Yoon (Principal, J. Lihm, V. Makarov, J. Degenhardt, J. O. Korbel (Principal, M. H. Fritz, S. Meiers, B. Raeder, T. Rausch, A. M. Stütz, P. Flicek (Principal, F. Paolo Casale, L. Clarke, R. E. Smith, O. Stegle, X. Zheng-Bradley, D. R. Bentley (Principal, B. Barnes, R. Keira Cheetham, M. Eberle, S. Humphray, S. Kahn, L. Murray, R. Shaw, E.-W. Lameijer, M. A. Batzer (Principal, M. K. Konkel, J. A. Walker, L. Ding (Principal, I. Hall, K. Ye, P. Lacroute, C. Lee (Principal, E. Cerveira, A. Malhotra, J. Hwang, D. Plewczynski, K. Radew, M. Romanovitch, C. Zhang, D. W. Craig (Principal, N. Homer, D. Church, C. Xiao, J. Sebat (Principal, D. Antaki, V. Bafna, J. Michaelson, K. Ye, S. E. Devine (Principal, E. J. Gardner (Project L, G. R. Abecasis (Principal, J. M. Kidd (Principal, R. E. Mills (Principal, G. Dayama, S. Emery, G. Jun, X. Shi (Principal, A. Quitadamo, G. Lunter (Principal, G. A. McVean (Principal, K. Chen (Principle, X. Fan, Z. Chong, T. Chen, D. Witherspoon, J. Xing, E. E. Eichler (Principal, M. J. Chaisson, F. Hormozdiari, J. Huddleston, M. Malig, B. J. Nelson, P. H. Sudmant, N. F. Parrish, E. Khurana (Principal, M. E. Hurles (Principal, B. Blackburne, S. J. Lindsay, Z. Ning, K. Walter, Y. Zhang, M. B. Gerstein (Principal, A. Abyzov, J. Chen, D. Clarke, H. Lam, X. Jasmine Mu, C. Sisu, J. Zhang, Y. Zhang, R. A. Gibbs (Principal, F. Yu (Project L, M. Bainbridge, D. Challis, U. S. Evani, C. Kovar, J. Lu, D. Muzny, U. Nagaswamy, J. G. Reid, A. Sabo, J. Yu, X. Guo, W. Li, Y. Li, R. Wu, G. T. Marth (Principal, E. P. Garrison, W. Fung Leong, A. N. Ward, G. del Angel, M. A. DePristo, S. B. Gabriel, N. Gupta, C. Hartl, R. E. Poplin, A. G.

Clark (Principal, J. L. Rodriguez-Flores, P. Flicek (Principal, L. Clarke, R. E. Smith, X. Zheng-Bradley, D. G. MacArthur (Principal, E. R. Mardis (Principal, R. Fulton, D. C. Koboldt, S. Gravel, C. D. Bustamante (Principal, D. W. Craig (Principal, A. Christoforides, N. Homer, T. Izatt, S. T. Sherry (Principal, C. Xiao, E. T. Dermitzakis (Principal, G. R. Abecasis (Principal, H. Min Kang, G. A. McVean (Principal, M. B. Gerstein (Principal, S. Balasubramanian, L. Habegger, H. Yu (Principal, P. Flicek (Principal, L. Clarke, F. Cunningham, I. Dunham, D. Zerbino, X. Zheng-Bradley, K. Lage (Principal, J. Berg Jespersen, H. Horn, S. B. Montgomery (Principal, M. K. DeGorter, E. Khurana (Principal, C. Tyler-Smith (Principal, Y. Chen, V. Colonna, Y. Xue, M. B. Gerstein (Principal, S. Balasubramanian, Y. Fu, D. Kim, A. Auton (Principal, A. Marcketta, R. Desalle, A. Narechania, M. A. Wilson Sayres, E. P. Garrison, R. E. Handsaker, S. Kashin, S. A. McCarroll, J. L. Rodriguez-Flores, P. Flicek (Principal, L. Clarke, X. Zheng-Bradley, Y. Erlich, M. Gymrek, T. Frederick Willems, C. D. Bustamante (Principal, F. L. Mendez, G. David Poznik, P. A. Underhill, C. Lee, E. Cerveira, A. Malhotra, M. Romanovitch, C. Zhang, G. R. Abecasis (Principal, L. Coin (Principal, H. Shao, D. Mittelman, C. Tyler-Smith (Principal, Q. Ayub, R. Banerjee, M. Cerezo, Y. Chen, T. W. Fitzgerald, S. Louzada, A. Massaia, S. McCarthy, G. R. Ritchie, Y. Xue, F. Yang, R. A. Gibbs (Principal, C. Kovar, D. Kalra, W. Hale, D. Muzny, J. G. Reid, J. Wang (Principal, X. Dan, X. Guo, G. Li, Y. Li, C. Ye, X. Zheng, D. M. Altshuler, P. Flicek (Principal, L. Clarke (Project L, X. Zheng-Bradley, D. R. Bentley (Principal, A. Cox, S. Humphray, S. Kahn, R. Sudbrak (Project L, M. W. Albrecht, M. Lienhard, D. Larson, D. W. Craig (Principal, T. Izatt, A. A. Kurdoglu, S. T. Sherry (Principal, C. Xiao, D. Haussler (Principal, G. R. Abecasis (Principal, G. A. McVean (Principal, R. M. Durbin (Principal, S. Balasubramaniam, T. M. Keane, S. McCarthy, J. Stalker, A. Chakravarti (Co-Chair), B. M. Knoppers (Co-Chair), G. R. Abecasis, K. C. Barnes, C. Beiswanger, E. G. Burchard, C. D. Bustamante, H. Cai, H. Cao, R. M. Durbin, N. P. Gerry, N. Gharani, R. A. Gibbs, C. R. Gignoux, S. Gravel, B. Henn, D. Jones, L. Jorde, J. S. Kaye, A. Keinan, A. Kent, A. Kerasidou, Y. Li, R. Mathias, G. A. McVean, A. Moreno-Estrada, P. N. Ossorio, M. Parker, A. M. Resch, C. N. Rotimi, Charmaine D. Royal, K. Sandoval, Y. Su, R. Sudbrak, Z. Tian, S. Tishkoff, L. H. Toji, C. Tyler-Smith, M. Via, Y. Wang, H. Yang, L. Yang, J. Zhu, W. Bodmer, G. Bedoya, A. Ruiz-Linares, Z. Cai, Y. Gao, J. Chu, L. Peltonen, A. Garcia-Montero, A. Orfao, J. Dutil, J. C. Martinez-Cruzado, T. K. Oleksyk, K. C. Barnes, R. A. Mathias, A. Hennis, H. Watson, C. McKenzie, F. Qadri, R. LaRocque, P. C. Sabeti, J. Zhu, X. Deng, P. C. Sabeti, D. Asogun, O. Folarin, C. Happi, O. Omoniwa, M. Stremlau, R. Tariyal, M. Jallow, F. Sisay Joof, T. Corrah, K. Rockett, D. Kwiatkowski, J. Kooner, T. Tịnh Hiê`n, S. J. Dunstan, N. Thuy Hang, R. Fonnie, R. Garry, L. Kanneh, L. Moses, P. C. Sabeti, J. Schieffelin, D. S. Grant, C. Gallo, G. Poletti, D. Saleheen, A. Rasheed, L. D. Brooks, A. L. Felsenfeld, J. E. McEwen, Y. Vaydylevich, E. D. Green, A. Duncanson, M. Dunn, J. A. Schloss, J. Wang, H. Yang, A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison, H. Min Kang, J. O. Korbel, J. L. Marchini, S. McCarthy, G. A. McVean, G. R. Abecasis, Corresponding authors, Steering committee, Production group, Baylor College of Medicine, BGI-Shenzhen, Broad Institute of MIT and Harvard, Coriell Institute for Medical Research, European Molecular Biology Laboratory, European Bioinformatics Institute, Illumina, Max Planck Institute for Molecular Genetics, McDonnell Genome Institute at Washington University, US National Institutes of Health, University of Oxford, Wellcome Trust Sanger Institute, Analysis group, Affymetrix, Albert Einstein College of Medicine, Baylor College of Medicine, BGI-Shenzhen, Bilkent University, Boston

College, Broad Institute of MIT and Harvard, Cold Spring Harbor Laboratory, Cornell University, European Molecular Biology Laboratory, European Molecular Biology Laboratory, European Bioinformatics Institute, Harvard University, Human Gene Mutation Database, Illumina, Icahn School of Medicine at Mount Sinai, Louisiana State University, Massachusetts General Hospital, Max Planck Institute for Molecular Genetics, McDonnell Genome Institute at Washington University, McGill University, National Eye Institute, NIH, New York Genome Center, Ontario Institute for Cancer Research, Pennsylvania State University, Rutgers Cancer Institute of New Jersey, Stanford University, Tel-Aviv University, The Jackson Laboratory for Genomic Medicine, T. F. Scientific, Translational Genomics Research Institute, US National Institutes of Health, University of California, San Diego, University of California, San Francisco, University of California, Santa Cruz, University of Chicago, University College London, University of Geneva, University of Maryland School of Medicine, University of Michigan, University of Montréal, University of North Carolina at Chapel Hill, University of North Carolina at Charlotte, University of Oxford, University of Puerto Rico, University of Texas Health Sciences Center at Houston, University of Utah, University of Washington, Weill Cornell Medical College, Wellcome Trust Sanger Institute, Yale University, Structural variation group, BGI-Shenzhen, Bilkent University, Boston College, Broad Institute of MIT and Harvard, Cold Spring Harbor Laboratory, Cornell University, European Molecular Biology Laboratory, European Molecular Biology Laboratory, European Bioinformatics Institute, Illumina, Leiden University Medical Center, Louisiana State University, McDonnell Genome Institute at Washington University, Stanford University, The Jackson Laboratory for Genomic Medicine, Translational Genomics Research Institute, US National Institutes of Health, University of California, San Diego, University of Maryland School of Medicine, University of Michigan, University of North Carolina at Charlotte, University of Oxford, University of Texas MD Anderson Cancer Center, University of Utah, University of Washington, Vanderbilt University School of Medicine, Weill Cornell Medical College, Wellcome Trust Sanger Institute, Yale University, Exome group, Baylor College of Medicine, BGI-Shenzhen, Boston College, Broad Institute of MIT and Harvard, Cornell University, European Molecular Biology Laboratory, European Bioinformatics Institute, Massachusetts General Hospital, McDonnell Genome Institute at Washington University, McGill University, Stanford University, Translational Genomics Research Institute, US National Institutes of Health, University of Geneva, University of Michigan, University of Oxford, Yale University, Functional interpretation group, Cornell University, European Molecular Biology Laboratory, European Bioinformatics Institute, Harvard University, Stanford University, Weill Cornell Medical College, Wellcome Trust Sanger Institute, Yale University, Chromosome Y group, Albert Einstein College of Medicine, American Museum of Natural History, Arizona State University, Boston College, Broad Institute of MIT and Harvard, Cornell University, European Molecular Biology Laboratory, European Bioinformatics Institute, New York Genome Center, Stanford University, The Jackson Laboratory for Genomic Medicine, University of Michigan, University of Queensland, Virginia Bioinformatics Institute, Wellcome Trust Sanger Institute, Data coordination center group, Baylor College of Medicine, BGI-Shenzhen, Broad Institute of MIT and Harvard, European Molecular Biology Laboratory, European Bioinformatics Institute, Illumina, Max Planck Institute for Molecular Genetics, McDonnell Genome Institute at Washington University, Translational Genomics Research Institute, US National Institutes of Health, University of

California, Santa Cruz, University of Michigan, University of Oxford, Wellcome Trust Sanger Institute, Samples and ELSI group, Sample collection, British from England and Scotland (GBR), Colombians in Medellín, Colombia (CLM), Han Chinese South (CHS), Finnish in Finland (FIN), Iberian Populations in Spain (IBS), Puerto Ricans in Puerto Rico (PUR), African Caribbean in Barbados (ACB), Bengali in Bangladesh (BEB), Chinese Dai in Xishuangbanna, China (CDX), Esan in Nigeria (ESN), Gambian in Western Division - Mandinka (GWD), Indian Telugu in the UK (ITU) and Sri Lankan Tamil in the UK (STU), Kinh in Ho Chi Minh City, Vietnam (KHV), Mende in Sierra Leone (MSL), Peruvian in Lima, Peru (PEL), Punjabi in Lahore, Pakistan (PJL), Scientific management, Writing group, A global reference for human genetic variation. *Nature*. **526**, 68–74 (2015).

31. J. Z. Li, D. M. Absher, H. Tang, A. M. Southwick, A. M. Casto, S. Ramachandran, H. M. Cann, G. S. Barsh, M. Feldman, L. L. Cavalli-Sforza, R. M. Myers, Worldwide human relationships inferred from genome-wide patterns of variation. *Science*. **319**, 1100–1104 (2008).

32. O. Sartor, J. S. de Bono, Metastatic prostate cancer. *N. Engl. J. Med.* **378**, 645–657 (2018).

33. D. A. Barocas, R. Grubb 3rd, A. Black, D. F. Penson, J. H. Fowke, G. Andriole, E. D. Crawford, Association between race and follow-up diagnostic care after a positive prostate cancer screening test in the prostate, lung, colorectal, and ovarian cancer screening trial. *Cancer*. **119**, 2223–2229 (2013).

34. Y. Han, K. A. Rand, D. J. Hazelett, S. A. Ingles, R. A. Kittles, S. S. Strom, B. A. Rybicki, B. Nemesure, W. B. Isaacs, J. L. Stanford, W. Zheng, F. R. Schumacher, S. I. Berndt, Z. Wang, J. Xu, N. Rohland, D. Reich, A. Tandon, B. Pasaniuc, A. Allen, D. Quinque, S. Mallick, D. Notani, M. G. Rosenfeld, R. S. Jayani, S. Kolb, S. M. Gapstur, V. L. Stevens, C. A. Pettaway, E. D. Yeboah, Y. Tettey, R. B. Biritwum, A. A. Adjei, E. Tay, A. Truelove, S. Niwa, A. P. Chokkalingam, E. M. John, A. B. Murphy, L. B. Signorello, J. Carpten, M. C. Leske, S.-Y. Wu, A. J. M. Hennis, C. Neslund-Dudas, A. W. Hsing, L. Chu, P. J. Goodman, E. A. Klein, S. L. Zheng, J. S. Witte, G. Casey, A. Lubwama, L. C. Pooler, X. Sheng, G. A. Coetzee, M. B. Cook, S. J. Chanock, D. O. Stram, S. Watya, W. J. Blot, D. V. Conti, B. E. Henderson, C. A. Haiman, Prostate cancer susceptibility in men of African ancestry at 8q24. *J. Natl. Cancer Inst.* **108**, djv431 (2016).

35. B. A. Mahal, A. A. Aizer, D. R. Ziehr, A. S. Hyatt, J. D. Sammon, M. Schmid, T. K. Choueiri, J. C. Hu, C. J. Sweeney, C. J. Beard, A. V. D'Amico, N. E. Martin, S. P. Kim, Q.-D. Trinh, P. L. Nguyen, Trends in disparate treatment of African American men with localized prostate cancer across National Comprehensive Cancer Network risk groups. *Urology*. **84**, 386–392 (2014).

36. K. Yamoah, M. H. Johnson, V. Choeurng, F. A. Faisal, K. Yousefi, Z. Haddad, A. E. Ross, M. Alshalafa, R. Den, P. Lal, M. Feldman, A. P. Dicker, E. A. Klein, E. Davicioni, T. R. Rebbeck, E. M. Schaeffer, Novel biomarker signature that may predict aggressive disease in African American men with prostate cancer. *J. Clin. Oncol.* **33**, 2789–2796 (2015).

37. H. Zhang, E. M. Messing, L. B. Travis, O. Hyrien, R. Chen, M. T. Milano, Y. Chen, Age and racial differences among PSA-detected (AJCC stage T1cN0M0) prostate cancer in the U.s.: A population-based study of 70,345 men. *Front. Oncol.* **3**, 312 (2013).

38. K. A. Moses, P. C. Sprenkle, G. Box, S. V. Carlsson, P. R. Carroll, W. J. Catalona, D. M. Dahl, NCCN Guidelines Prostate Cancer Early Detection Version 1.2022 (2021).

39. D. A. Siegel, M. E. O'Neil, T. B. Richards, N. F. Dowling, H. K. Weir, Prostate cancer incidence and survival, by stage and race/ethnicity - United States, 2001-2017. *MMWR Morb. Mortal. Wkly. Rep.* **69**, 1473–1480 (2020).

40. M. B. Cook, P. S. Rosenberg, F. A. McCarty, M. Wu, J. King, C. Eheman, W. F. Anderson, Racial disparities in prostate cancer incidence rates by census division in the United States, 1999-2008. *Prostate*. **75**, 758–763 (2015).

41. Atlas of cancer mortality in the United States, 1950-94. *Int. J. Epidemiol.* **29**, 602-a-602 (2000).

42. A. Jemal, E. Ward, X. Wu, H. J. Martin, C. C. McLaughlin, M. J. Thun, Geographic patterns of prostate cancer mortality and variations in access to medical care in the United States. *Cancer Epidemiol. Biomarkers Prev.* **14**, 590–595 (2005).

43. S. J. Curry, A. H. Krist, D. K. Owens, Annual report to the nation on the status of cancer, part II: Recent changes in prostate cancer trends and disease characteristics. *Cancer*. **125** (2019), pp. 317–318.

44. D. Marnetto, K. Pärna, K. Läll, L. Molinaro, F. Montinaro, T. Haller, M. Metspalu, R. Mägi, K. Fischer, L. Pagani, Ancestry deconvolution and partial polygenic score can improve susceptibility predictions in recently admixed individuals. *Nat. Commun.* **11**, 1628 (2020).

45. I. J. Powell, The precise role of ethnicity and family history on aggressive prostate cancer: a review analysis. *Arch. Esp. Urol.* **64**, 711–719 (2011).

46. A. M. D. Wolf, R. C. Wender, R. B. Etzioni, I. M. Thompson, A. V. D'Amico, R. J. Volk, D. D. Brooks, C. Dash, I. Guessous, K. Andrews, C. DeSantis, R. A. Smith, American Cancer Society Prostate Cancer Advisory Committee, American Cancer Society guideline for the early detection of prostate cancer: update 2010. *CA Cancer J. Clin.* **60**, 70–98 (2010).

47. A. Horwich, J. Hugosson, T. de Reijke, T. Wiegel, K. Fizazi, V. Kataja, Panel Members, European Society for Medical Oncology, Prostate cancer: ESMO Consensus Conference Guidelines 2012. *Ann. Oncol.* **24**, 1141–1162 (2013).

48. A. Qaseem, M. J. Barry, T. D. Denberg, D. K. Owens, P. Shekelle, Clinical Guidelines Committee of the American College of Physicians, Screening for prostate cancer: a guidance statement from the Clinical Guidelines Committee of the American College of Physicians. *Ann. Intern. Med.* **158**, 761–769 (2013).

49. Z. Shi, E. A. Platz, J. Wei, R. Na, R. J. Fantus, C.-H. Wang, S. E. Eggener, P. J. Hulick, D. Duggan, S. L. Zheng, K. A. Cooney, W. B. Isaacs, B. T. Helfand, J. Xu, Performance of three inherited risk measures for predicting prostate cancer incidence and mortality: A population-based prospective analysis. *Eur. Urol.* **79**, 419–426 (2021).

50. L. Kachuri, R. E. Graff, K. Smith-Byrne, T. J. Meyers, S. R. Rashkin, E. Ziv, J. S. Witte, M. Johansson, Pan-cancer analysis demonstrates that integrating polygenic risk scores with modifiable risk factors improves risk prediction. *Nat. Commun.* **11**, 6084 (2020).

51. A. Schecter, L. Needham, M. Pavuk, J. Michalek, J. Colacino, J. Ryan, O. Päpke, L. Birnbaum, Agent Orange exposure, Vietnam war veterans, and the risk of prostate cancer. *Cancer*. **115** (2009), pp. 3369–3371.

52. A. Plym, K. L. Penney, S. Kalia, P. Kraft, D. V. Conti, C. Haiman, L. A. Mucci, A. S. Kibel, Evaluation of a multiethnic polygenic risk score model for prostate cancer. *J. Natl. Cancer Inst.* (2021), doi:10.1093/jnci/djab058.

53. M. Eklund, F. Jäderling, A. Discacciati, M. Bergman, M. Annerstedt, M. Aly, A. Glaessgen, S. Carlsson, H. Grönberg, T. Nordström, STHLM3 consortium, MRI-targeted or standard biopsy in prostate cancer screening. *N. Engl. J. Med.* **385**, 908–920 (2021).

54. Y. Jiang, T. J. Meyers, A. A. Emeka, L. F. Cooley, P. R. Cooper, N. Lancki, I. Helenowski, L. Kachuri, D. W. Lin, J. L. Stanford, L. F. Newcomb, S. Kolb, A. Finelli, N. E. Fleshner, M. Komisarenko, J. A. Eastham, B. Ehdaie, N. Benfante, C. J. Logothetis, J. R. Gregg, C. A. Perez, S. Garza, J. Kim, L. S. Marks, M. Delfin, D. Barsa, D. Vesprini, L. H. Klotz, A. Loblaw, A. Mamedov, S. L. Goldenberg, C. S. Higano, M. Spillane, E. Wu, H. B. Carter, C. P. Pavlovich, M. Mamawala, T. Landis, P. R. Carroll, J. M. Chan, M. R. Cooperberg, J. E. Cowan, T. M. Morgan, J. Siddiqui, R. Martin, E. A. Klein, K. Brittain, P. Gotwald, D. A. Barocas, J. R. Dallmer, J. B. Gordetsky, P. Steele, S. D. Kundu, J. Stockdale, M. J. Roobol, L. D. F. Venderbos, M. G. Sanda, R. Arnold, D. Patil, C. P. Evans, M. A. Dall'Era, A. Vij, A. J. Costello, K. Chow, N. M. Corcoran, S. Rais-Bahrami, C. Phares, D. S. Scherr, T. Flynn, R. J. Karnes, M. Koch, C. R. Dhondt, J. B. Nelson, D. McBride, M. S. Cookson, K. L. Stratton, S. Farriester, E. Hemken, W. M. Stadler, T. Pera, D. Banionyte, F. J. Bianco Jr, I. H. Lopez, S. Loeb, S. S. Taneja, N. Byrne, C. L. Amling, A. Martinez, L. Boileau, F. D. Gaylis, J. Petkewicz, N. Kirwen, B. T. Helfand, J. Xu, D. M. Scholtens, W. J. Catalona, J. S. Witte, Genetic factors associated with prostate cancer conversion from active surveillance to treatment. *HGG Adv*. **3**, 100070 (2022).

55. A. J. Vickers, D. Ulmert, D. D. Sjoberg, C. J. Bennette, T. Björk, A. Gerdtsson, J. Manjer, P. M. Nilsson, A. Dahlin, A. Bjartell, P. T. Scardino, H. Lilja, Strategy for detection of prostate cancer based on relation between prostate specific antigen at age 40-55 and long term risk of metastasis: case-control study. *BMJ*. **346**, f2023 (2013).

56. H. Lilja, D. Ulmert, T. Björk, C. Becker, A. M. Serio, J.-A. Nilsson, P.-A. Abrahamsson, A. J. Vickers, G. Berglund, Long-term prediction of prostate cancer up to 25 years before diagnosis of prostate cancer using prostate kallikreins measured at age 44 to 50 years. *J. Clin. Oncol.* **25**, 431–436 (2007).

57. US Preventive Services Task Force, D. C. Grossman, S. J. Curry, D. K. Owens, K. Bibbins-Domingo, A. B. Caughey, K. W. Davidson, C. A. Doubeni, M. Ebell, J. W. Epling Jr, A. R. Kemper, A. H. Krist, M. Kubik, C. S. Landefeld, C. M. Mangione, M. Silverstein, M. A. Simon, A. L. Siu, C.-W. Tseng, Screening for prostate cancer. *JAMA*. **319**, 1901 (2018).

58. A. Plym, Y. Zhang, K. Stopsack, B. Delcoigne, A. S. Kibel, E. Giovannucci, K. L. Penney, L. A. Mucci, in *Epidemiology* (American Association for Cancer Research, 2021; http://dx.doi.org/10.1158/1538-7445.am2021-822).

59. M. Ahmed, C. Goh, E. Saunders, C. Cieza-Borrella, PRACTICAL consortium, Z. Kote-Jarai, F. R. Schumacher, R. Eeles, Germline genetic variation in prostate susceptibility does not predict outcomes in the chemoprevention trials PCPT and SELECT. *Prostate Cancer Prostatic Dis.* **23**, 333–342 (2020).