Title

Genetic mapping, inference and prediction across diverse human populations

Author

Hou, Kangcheng

Publication Date

2024

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Genetic mapping, inference and prediction across diverse human populations

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

by

Kangcheng Hou

2024

ABSTRACT OF THE DISSERTATION

Genetic mapping, inference and prediction

across diverse human populations

by

Kangcheng Hou

Doctor of Philosophy in Bioinformatics

University of California, Los Angeles, 2024

Professor Bogdan Pasaniuc, Chair

Genome-wide association studies have revolutionized our understanding of genetic influences on common diseases and complex traits. However, the majority of discoveries have been limited to individuals of European ancestry, leading to a data collection bias that disproportionately under-samples non-European populations. This bias leads to missed discovery opportunities and differential prediction accuracy across sub-populations defined by genetic ancestry and socioeconomic factors. Although datasets with diverse genetic ancestry backgrounds are increasingly available, existing analytical tools often fail to account for the heterogeneity present

in these datasets. Here, I introduce new computational and statistical methods for genetic mapping, inference, and prediction across diverse human populations. First, I investigate the power of genetic mapping approaches in populations with diverse genetic ancestry backgrounds. Second, I explore the inference of genetic architecture, estimating the cross-ancestry sharing of genetic effects. Third, I examine genetic prediction, quantifying differential polygenic scoring accuracy by contexts and developing an approach to account for such differences.

The dissertation of Kangcheng Hou is approved.

Kirk Edward Lohmueller

Michael Jeffrey Gandal

Sriram Sankararaman

Bogdan Pasaniuc, Committee Chair

University of California, Los Angeles

2024

This dissertation is dedicated to my parents.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First, I want to thank my PhD advisor, Bogdan Pasaniuc. Bogdan introduced me to the field of statistical genetics and guided me with his extensive knowledge. He taught me to identify impactful problem formulations where I can make unique contributions. He provided detailed hands-on feedback and allowed me the freedom to explore as I needed them. His continuous support has shaped me into a better researcher. Bogdan has set a perfect example of a great mentor that I want to be.

I am fortunate to have Kirk Lohmueller, Michael Gandal, and Sriram Sankararaman on my committee. Kirk's work has been my go-to resource for learning more about population genetics. Mike's enthusiasm for biology of psychiatric diseases motivates me to dive into my analyses to make meaningful contribution to understand disease biology. Sriram's guidance and feedback have been crucial to the statistical and computational aspects of my work.

I am grateful for the opportunities to collaborate with so many exceptional researchers. Huwenbo Shi and Kathy Burch got me started on statistical genetics research. Huwenbo taught me all the math about heritability in his and other's papers and guided me in writing my first program for analyzing genetics data. Kathy taught me how to derive rigorous math formulations for new research problems, and the importance of effective presentations, both in writing concise well-structured texts and creating intuitive thoughtful figures. Yi Ding and I joined the lab around the same time; we grew and collaborated on many projects together. I appreciate the countless hours we spent brainstorming ideas, troubleshooting challenges, and celebrating every small success along the way. These experiences have been among the most rewarding and meaningful moments of my graduate school. Ziqi Xu worked with me when she was an undergraduate student

at UCLA; she taught me about being curious and enthusiastic about research. I also had the fortune to work with Martin Zhang and Alkes Price. Martin taught me the importance of rigorous statistical analysis and the courage to take challenging but important research questions. Alkes taught me the power of simplicity and consistency. I want to thank all of my collaborators for their contributions for the work presented in this dissertation, including Elizabeth Atkinson, Steven Gazal, Stephanie Gogarten, Buhm Han, Arbel Harpak, Matthew Heffel, Jibril Hirbo, Joohyun Kim, Hanbin Lee, Xihao Li, Yun Li, Chongyuan Luo, Alicia Martin, Ashok Patowary, Florian Privé, Jonathan Shortt, Taotao Tan, Bjarni Vilhjálmsson, Xinan Wang, Ying Wang, Haoyu Zhang, Jingtian Zhou. I always learn something valuable from each interaction with them.

I am lucky to be in such a supportive environment within UCLA, with lab mates, colleagues, faculties and staffs. They include Nick Bayley, Toni Boltz, Kristin Boulier, Leah Briscoe, Alec Chiu, Chenlu Di, Alex Flynn-Carroll, Malika Kumar Freund, Boyang Fu, Claudia Giambartolomei, Claudia Giambartolomei, Jonatan Hervoso, Ruth Johnson, Sergey Knyazev, Sandra Lapinska, Xinzhe Li, Choo Liu, Megan Major, Arun Majumdar, Nicholas Mancuso, Ravi Mandla, Igor Mandric, Rachel Mester, Ali Pazokitoroudi, Ella Petter, Harold Pimentel, Tommer Schwarz, Zhuozheng Shi, Cyrillus Tan, Veronica Tozzo, Dongyuan Song, Vidhya Venkaterasewan, Ariel Wu, Yang Wu, Yu Yan. They have consistently shown me kindness and warmth. Special thanks to Gene Gray, Grace Xiao, Gayane Hovhannisyan in overseeing and administering the Bioinformatics and Computational Medicine program so efficiently.

I especially thank my partner. Her support, passion, and wisdom has made these years all the more meaningful and joyful. I treasure all the moments, the highs and lows, we have been together during our graduate school and beyond.

Lastly, I thank my family for their selfless and continued support. My older brother, Kanglei Fang has been setting an example for me to follow, both personally and professionally. I thank my parents Min Fang and Zhijiang Hou, for providing me with the best education and the opportunities that have shaped me as who I am today. They have always believed in me and unconditionally supported me in every step of my life. I dedicate this thesis to them.

Chapter 2 is published in Nature Genetics as Kangcheng Hou, Arjun Bhattacharya, Rachel Mester, Kathryn S. Burch, Bogdan Pasaniuc. On powerful GWAS in admixed populations. Nat Genet 53, 1631–1633 (2021). https://doi.org/10.1038/s41588-021-00953-5. K.H. and B.P. conceived and designed the experiments. K.H. performed the experiments and statistical analyses. K.H., K.S.B., R.M. and A.B. collected and managed the data. K.H., K.S.B., R.M., A.B. and B.P. wrote the manuscript.

Chapter 3 is published in Nature Genetics as Kangcheng Hou, Yi Ding, Ziqi Xu, Yue Wu, Arjun Bhattacharya, Rachel Mester, Gillian M. Belbin, Steve Buyske, David V. Conti, Burcu F. Darst, Myriam Fornage, Chris Gignoux, Xiuqing Guo, Christopher Haiman, Eimear E. Kenny, Michelle Kim, Charles Kooperberg, Leslie Lange, Ani Manichaikul, Kari E. North, Ulrike Peters, Laura J. Rasmussen-Torvik, Stephen S. Rich, Jerome I. Rotter, Heather E. Wheeler, Genevieve L. Wojcik, Ying Zhou, Sriram Sankararaman & Bogdan Pasaniuc. Causal effects on complex traits are similar for common variants across segments of different continental ancestries within admixed individuals. Nat Genet 55, 549–558 (2023). https://doi.org/10.1038/s41588-023-01338-6. K.H. and B.P. conceived and designed the experiments. K.H. performed the experiments and statistical analyses with assistance from Y.D., Z.X., Y.W., A.B., R.M., S.S. and B.P. G.M.B., S.B., D.V.C., B.F.D., M.F., C.G., X.G., C.H., E.E.K., M.K., C.K., L.L., A.M., K.E.N., U.P., L.J.R.-T., S.S.R., J.I.R.,

H.E.W., G.L.W. and Y.Z. provided data and feedback on analysis. K.H. and B.P. wrote the manuscript with feedback from all authors.

Chapter 4 is in-press in Nature Genetics as Kangcheng Hou, Ziqi Xu, Yi Ding, Ravi Mandla, Zhuozheng Shi, Kristin Boulier, Arbel Harpak, Bogdan Pasaniuc. Calibrated prediction intervals for polygenic scores across diverse contexts. A preprint version is available https://doi.org/10.1101/2023.07.24.23293056. K.H. and B.P. conceived and designed the experiments. K.H., Z.X. and Y.D. performed the experiments and statistical analyses with assistance from R.M., Z.S., K.B., A.H. and B.P. K.H. and B.P. wrote the manuscript with feedback from all authors.

VITA

EDUCATION

2015-2019 B.Eng., Computer Science

Zhejiang University | Hangzhou, China

2019-2024 PhD candidate, Bioinformatics

University of California, Los Angeles | Los Angeles, CA

SELECTED PUBLICATIONS

[1] Kangcheng Hou, Ziqi Xu, Yi Ding, Arbel Harpak, and Bogdan Pasaniuc. **Calibrated prediction intervals for polygenic scores across diverse contexts**. *medRxiv*, 2023.

[2] Kangcheng Hou, Yi Ding, Ziqi Xu, Yue Wu, Arjun Bhattacharya, Rachel Mester, Gillian M Belbin, Steve Buyske, David V Conti, Burcu F Darst, Myriam Fornage, Chris Gignoux, Xiuqing Guo, Christopher Haiman, Eimear E Kenny, Michelle Kim, Charles Kooperberg, Leslie Lange, Ani Manichaikul, Kari E North, Ulrike Peters, Laura J Rasmussen-Torvik, Stephen S Rich, Jerome I Rotter, Heather E Wheeler, Genevieve L Wojcik, Ying Zhou, Sriram Sankararaman, and Bogdan Pasaniuc. **Causal effects on complex traits are similar for common variants across segments of different continental ancestries within admixed individuals**. *Nature Genetics*, 2023.

[3] Yi Ding, Kangcheng Hou, Ziqi Xu, Aditya Pimplaskar, Ella Petter, Kristin Boulier, Florian Privé, Bjarni J Vilhjálmsson, Loes M Olde Loohuis, and Bogdan Pasaniuc. **Polygenic scoring accuracy varies across the genetic ancestry continuum**. *Nature*, 2023.

[4] Martin Jinye Zhang, Kangcheng Hou, Kushal K Dey, Saori Sakaue, Karthik A Jagadeesh, Kathryn Weinand, Aris Taychameeki- atchai, Poorvi Rao, Angela Oliveira Pisco, James Zou, Bruce Wang, Michael Gandal, Soumya Raychaudhuri, Bogdan Pasaniuc, and Alkes L Price. **Polygenic enrichment distinguishes disease associations of individual cells in single-cell RNA-seq data**. *Nature Genetics*, 2022.

[5] Yi Ding, Kangcheng Hou, Kathryn S Burch, Sandra Lapinska, Florian Privé, Bjarni Vilhjálmsson, Sriram Sankararaman, and Bogdan Pasaniuc. **Large uncertainty in individual polygenic risk score estimation**

impacts PRS-based risk stratification. *Nature Genetics*, 2022.

[6] Kathryn S Burch, Kangcheng Hou, Yi Ding, Yifei Wang, Steven Gazal, Huwenbo Shi, and Bogdan Pasaniuc. **Partitioning gene- level contributions to complex-trait heritability by allele frequency identifies disease-relevant genes**. *The American Journal of Human Genetics*, 2022.

[7] Kangcheng Hou, Arjun Bhattacharya, Rachel Mester, Kathryn S Burch, and Bogdan Pasaniuc. **On powerful GWAS in admixed populations**. *Nature Genetics*, 2021.

[8] Kangcheng Hou, Kathryn S Burch, Arunabha Majumdar, Huwenbo Shi, Nicholas Mancuso, Yue Wu, Sriram Sankararaman, and Bogdan Pasaniuc. **Accurate estimation of SNP-heritability from biobank-scale data irrespective of genetic architecture**. *Nature Genetics*, 2019.

[9] Kangcheng Hou, Stephanie Gogarten, Joohyun Kim, Xing Hua, Julie-Alexia Dias, Quan Sun, Ying Wang, Taotao Tan, Elizabeth G Atkinson, Alicia Martin, Jonathan Shortt, Jibril Hirbo, Yun Li, Bogdan Pasaniuc, and Haoyu Zhang. **Admix-kit: an integrated toolkit and pipeline for genetic analyses of admixed populations**. *Bioinformatics*, 2024.

1 Introduction

Complex traits and common diseases are influenced by both genetic and environment factors[1,2]. While environments change over time and are challenging to comprehensively measure, an individual's genetic information is largely unchanged throughout lifetime. With the substantially decreasing DNA sequencing cost[3], integration of genetic information is becoming a cost-effective approach towards precision medicine, tailoring medical prevention and treatment according to unique profile of each individual[4].

Genome-wide association studies (GWAS), a representative approach for genetic mapping that links candidate genetic factors to disease status, have been fruitful over the past two decades[5–7]. For example, a CRISPR-based drug, Casgevy, was recently approved by FDA as the first gene therapy for treating sickle cell disease – the corresponding drug target gene *BCL11A* was initially discovered in GWAS of fetal hemoglobin levels dating back to 2007[8,9]. Broadly, the success rate of drug targets with genetic support from GWAS is more than two times greater than those without[10,11]. In addition to prioritizing individual disease-associated genetic variants, polygenic scores aggregating effects of multiple disease-associated genetic variants, have become a powerful tool to predict disease risk[12], inform groups of individuals most likely to benefit from treatment[13], as well improving clinical practice, including adjustment for genetic component of lab value[14]. The use of genetic data will continue to inform drug development, and change how we practice medicine incorporating personalized information.

The above achievements have largely been established within individuals of European ancestry due to historical data collection bias disproportionally under-sampling individuals of non-European ancestry[15]. As GWAS have detection power for variants with sufficient amount of variation within

a sample, such bias is missing important discovery opportunities[16]. For example, genetic variants in gene *SLC16A11* are rare in European but common in Native American and East Asian ancestry backgrounds. And their link to type 2 diabetes would be otherwise missed if GWAS were performed in Europeans only[17]. Moreover, existing polygenic scoring methods have differential prediction accuracy across ancestry groups[18–20] – disease risk predicted by such models are not as accurate in other ancestry groups compared to Europeans. Other than genetic ancestry, individual-level contexts including age, sex, or socioeconomic status also impact accuracy[21].

To realize the promise of precision medicine for everyone, the field has put significant efforts to collect genetic data samples that are representative of the world populations[16,22–24]. However, existing analytical tools often fail to capture the heterogeneity and diversity present in these datasets. Key questions remain regarding the computational and statistical methodologies for genetic mapping, inference and prediction across diverse populations. In **Chapter 2**, I study the power of genetic mapping approaches in populations with diverse genetic ancestry backgrounds (a version is published in *Nature Genetics*[25]). In **Chapter 3**, I study the inference of genetic architecture, quantifying the sharing of genetic effects across ancestry backgrounds (a version is published in *Nature Genetics*[26]). In **Chapter 4**, I study genetic prediction, where I quantify differential polygenic scoring accuracy by contexts and for which I develop an approach to account (a version is available as a preprint[27]). Central to these works is the development of new data-driven approach for modeling and accounting for the diversity of genetic ancestry and context backgrounds across human populations.

2 On powerful GWAS in admixed populations

2.1 Introduction

Improving statistical power for GWAS in admixed populations is imperative, as more and larger genomic studies in admixed populations are desperately needed to accelerate genomic medicine and reduce health inequities[28]. Recently, Atkinson et al.[29] introduced a statistical framework (Tractor) for GWAS in admixed populations (e.g., African Americans) that corrects for population structure through the use of local ancestry and concluded that GWAS in admixed populations increases discovery power over traditional GWAS only in the presence of allelic effect-size heterogeneity by ancestry; a decrease in power is expected when allelic effects at tested variants are similar across ancestries. We wish to clarify that Atkinson et al.'s conclusion is specific to their particular choice of statistical association test that prioritizes allelic effect-size heterogeneity by ancestry and does not hold for other existing tests for GWAS in admixed populations. Existing association tests attain increased power over traditional GWAS in admixed populations, even when the causal variant has similar allelic effects across ancestries[30–32]. Therefore, GWAS in admixed populations increases power for discovery over homogeneous populations in either scenario—similar or different ancestry-specific allelic effects.

Powerful GWAS in admixed populations when causal variant has similar allelic effects across ancestries is performed either through explicit modeling of the relationships between allelic and local-ancestry effects[32–35] or implicit inclusion of the admixture signal in tests that do not correct for local ancestry[30,36]. In all approaches, population structure is appropriately controlled by correcting for global ancestry[30]. The gain in power stems from differentiation of causal-allele frequencies by ancestry that induces heterogeneity in the standardized ancestry-specific effects,

which in turn induces a local-ancestry effect on the trait. Therefore, larger power gains over traditional GWAS are expected for causal variants with higher degrees of frequency differentiation between ancestral populations[30,32]. Most importantly, by using such tests, GWAS in African American individuals attain superior power relative to GWAS in ancestrally homogeneous populations such as Europeans or Africans[30–32]. Therefore, when allelic effects are similar across ancestries, correcting for local ancestry is expected to impair statistical power for GWAS discovery as compared to global ancestry adjustment[36] and is more useful as a localization tool in post-GWAS fine-mapping[30].

2.2 Results

We use simulations to compare the test proposed by Atkinson et al. (Tractor) to existing methods for GWAS in admixed populations when causal allelic effects are similar across ancestries[31]. Starting from 1000 Genomes genotypes[37], we simulated 40,000 admixed individuals assuming admixture fractions of 80% African and 20% European followed by 7 generations of random mating (Figure 1A). We simulated a phenotype with 10% prevalence under the Tractor logistic model with a single causal variant with the same allelic effect across ancestries[29]; variability in causal variant frequencies across ancestries induce heterogeneity by ancestry in the marginal standardized effects. We compared the following tests for disease mapping in admixed populations: ATT (Armitage trend test with correction for global ancestry); ATT-Logit (logistic regression with genotypic effects only; this test is similar to that used by the PAGE study[36]); ADM (case-only admixture mapping); ADM-Logit (case-control admixture mapping; similar to the M1 model of Atkinson et al.); SNP1 (association conditioned on local ancestry; similar to the M2 model referred to as “traditional GWAS” in Atkinson et al.); MIX (combined case-only admixture and SNP case–control association)[32]; SUM (sum of case–control SNP association and case-only admixture

association); and Tractor (logistic regression assuming independent effects across ancestries with correction for local ancestry)[29]. All tests correct for global ancestry; SUM and Tractor are 2-degrees-of-freedom (dof) tests while all others are 1-dof tests.

First, we find that all tests appropriately control false positive rates under the null hypothesis (Supplementary Fig. 1). Second, as previously reported, we find that 1-dof methods that only correct for global ancestry (ATT, ATT-Logit, MIX) attain superior power over methods that correct for both global and local ancestry (SNP1/Tractor-M2). As expected, a larger gain in power is observed at SNPs with higher frequency differentiation by ancestry. Since SNP1 and Tractor-M2 are analogous to disease mapping in ancestrally homogeneous populations (see refs[30,32]), it follows that admixed populations can offer increased power for disease mapping as compared to ancestrally homogeneous populations. For example, when OR=1.2 for a causal variant uniformly drawn from the genome, in a GWAS of 4,000 cases and 4,000 controls, ATT and MIX yield ~27% power compared to 25% for SNP1/Tractor-M2 and 20% for Tractor (Figure 1A). A larger gain in power is observed at causal variants with frequency differentiation greater than 0.2 between ancestries (28% of all variants), where we observe a power of 43% for MIX, 33% for SNP1/Tractor-M2, and 26% for Tractor (Figure 1A). Tractor has reduced power in these simulations as it requires some degree of heterogeneity in allelic effects to improve power (e.g., more than 60% difference in allelic effects when frequency is fixed across ancestries[29]). Similar results were observed at other effect sizes or when the causal variant is untyped and missing from the data, thus confirming that GWAS in admixed populations outperforms traditional GWAS when the causal variant has similar allelic effects across ancestries (Figure 1A, Supplementary Figs. 2-3).

Next we analyzed GWAS data of real lipid phenotypes - total cholesterol (TC) and low-density lipoprotein cholesterol (LDL) - in individuals of African–European ancestries within UK Biobank (N=4,327, Data availability). We focused on four well-known regions containing GWAS signals for lipid traits (*LDLR*, *APOE*, *PCSK9*, *SORT1)*. Similar to simulations, we observe that the association with correction for genome-wide ancestry only (ATT) yields the strongest signal, followed by tests that correct for both local and global ancestry (SNP1). Tractor, which also models heterogeneous effects, yielded the weakest association signal (Table 1). For example, at the *LDLR* region ATT attains P = 2.3e-10 followed by 2.76e-10 for SNP1 and 1.64e-09 for Tractor (Figure 1B). Notably, averaging across the four regions, Tractor yields ~11% decreased effective sample size compared to ATT. For an extensive evaluation of admixture-aware tests at risk regions under strong admixture peaks, we refer to ref[32].

2.3 Discussion

In conclusion, GWAS in admixed populations attain improved power for discovery over homogeneous populations in either scenario—similar or different ancestry-specific allelic effects—thus further supporting the need for larger genomic studies in such populations. Here, we show that disease mapping in admixed populations is well powered when allelic effects are similar across ancestries, whereas Atkinson et al. showcase the power gains from 2-dof tests in the presence of effect-size heterogeneity by ancestry[29,30,32]. Since the true extent of heterogeneity in causal allelic effects across ancestries is currently unknown[38–42], we recommend careful consideration of the balance between expected allelic effect-size heterogeneity across ancestries and association power when selecting a statistical test for GWAS in admixed populations. A further consideration should be given to linkage-disequilibrium induced heterogeneity at tagging variants which occurs even when causal allelic effects are similar across ancestries[29,30,32]; in this

scenario there is an expected loss of power due to imperfect tagging, although preliminary results suggest that the loss in power is small particularly when genotype imputation is employed (Supplementary Fig. 3, also see Table 2 of ref[30,32]). Properly aligned statistical tests will enable novel discoveries in admixed populations that have long been understudied and underserved.

2.4 Figures

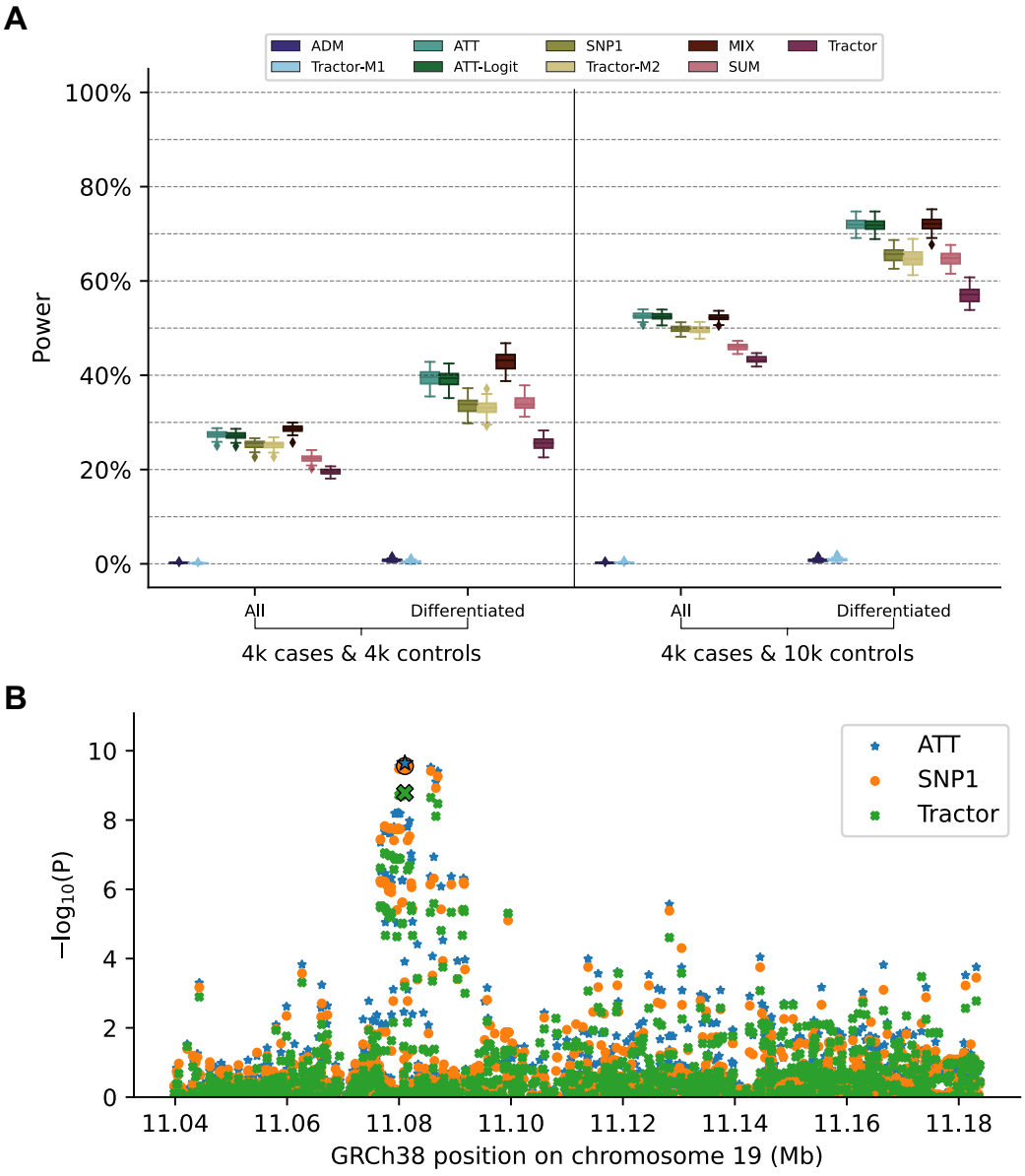

Figure 2.1 Power of GWAS tests in simulations and in real data.

(A) Comparison of power of GWAS tests in admixed populations in simulations. The boxplots denoted “All” represent distributions of power estimates from 50 simulation replicates and 3,000

causal SNPs uniformly drawn from the set of all SNPs (150,000 points per boxplot), while "Differentiated" restricts to the subset of SNPs (904 out of 3,000) with absolute allele frequency difference > 0.2 between Europeans and Africans (45,200 points per boxplot). For box plots, the central lines correspond to the medians. The boxes represent the first and third quartiles of the points. Whiskers represent the minimum and maximum points located within 1.5 × interquartile range (IQR) from the first and third quartiles, respectively. Here, we present results for an odds ratio (OR) of 1.2; additional results, including null simulations, can be found at Supplementary Fig. 2. **(B)** $-\log_{10}(p)$ of SNP associations with LDL in the *LDLR* locus. The SNP with the strongest *Tractor* association p-value is framed and enlarged. Results at other considered GWAS regions for lipids (*APOE*, *PCSK9*, *SORT1*) show similar patterns (Table 1).

2.5 Tables

Trait	**Locus**	**ATT**	**SNP1**	**Tractor**
TC	*APOE*	30.6	30.3 (-1.0%)	28.9 (-5.6%)
LDL	*APOE*	50	49.8 (-0.5%)	47.5 (-5.1%)
TC	*LDLR*	8.3	8.2 (-0.9%)	7.6 (-8.4%)
LDL	*LDLR*	9.6	9.6 (-0.8%)	8.8 (-8.9%)

TC	*PCSK9*	9.4	8.5 (-9.9%)	7.7 (-18.3%)
LDL	*PCSK9*	9.6	9.4 (-1.3%)	8.5 (-10.9%)
TC	*SORT1*	5.1	5.0 (-0.9%)	4.3 (-15.7%)
LDL	*SORT1*	7.1	7.1 (-0.5%)	6.3 (-11.7%)
Average relative difference			**-2.0%**	**-10.6%**

Table 2.1 -log10 P-values association statistics for the top Tractor SNP at known risk loci.

We considered three GWAS tests with correction for: global ancestry (*ATT*); global and local ancestry (*SNP1*); global and local ancestry while allowing for heterogeneous effects (*Tractor*). Index SNP was selected based on the strongest Tractor association p-value. Relative differences to the ATT score are shown in parentheses and the last line.

3 Causal effects on complex traits are similar for common variants across segments of different continental ancestries within admixed individuals

3.1 Introduction

Large-scale genotype-phenotype studies are increasingly analyzing diverse sets of individuals of various continental and sub-continental ancestries[16,22,43,44]. A fundamental open question in these studies is to what extent the genetic basis of common human diseases and traits are shared/distinct across different ancestry populations and its impact to genetic discovery and prediction[45–49]. For example, it is unclear how much of the low polygenic score portability can be attributed to differences in genetic causal effects across ancestries[18,45,50]. Hence, understanding the role of ancestry in variability of causal effect sizes has tremendous implications for understanding the genetic basis of disease and portability of genetic risk scores in personalized and equitable genomic medicine[16,18,50–52].

The standard approach to estimating similarity in causal effects across ancestries has focused on cross-population analyses (typically at continental level) in which effect sizes estimated by large-scale genome-wide association studies (GWAS) are compared across continental-level ancestry groups[45–48,53,54]. Such studies have found significant differences, albeit with modest magnitude, of causal effects in cross-continental comparisons. However, a main drawback of such studies is the differences in definition of environment/phenotype across such broad units of ancestry that can reduce the observed similarity; for example, the low estimated similarity in causal genetic

effects for Major Depressive Disorder across Europeans and East Asians may be attributed to different diagnostic criteria in the two populations[48,55].

As an alternative to studying populations across different continents, causal effects similarity by ancestry can also be studied within recently admixed populations. Recently admixed individuals have the unique feature of having their genomes as mosaic of ancestry segments (*local ancestry*) originating from the ancestral populations within the past few dozen generations; for example, African American genomes are comprised of segments of African and European ancestries within the past 5-15 generations[31]. Unfortunately, admixed populations are vastly under-represented in genomic studies[15], partly because of the lack of understanding of how the genetic causal effects vary across ancestries[25,29,31,50,56,57]. For example, heterogeneity of marginal effects (which is estimated in GWAS single variant scan and can tag effects from nearby variants due to linkage disequilibrium (LD)) for a few traits and loci has been reported[58–61], but it remains unknown whether this reflects true difference in causal genetic effects or confounding due to different allele frequencies and/or LD by ancestry. Recent work[54] have reported evidence of causal effect heterogeneity for SNPs in regions of European ancestries comparing individuals of European versus African American ancestries; however, these studies focused on cross-population comparisons instead of comparing effects across local ancestries within admixed populations. Estimating the magnitude of similarity in causal effects across ancestries is important for all genotype-phenotype studies in admixed populations from mapping to polygenic prediction, particularly within methods that allow for effects to vary across local ancestry segments[25,29,56,57].

In this work, we quantify the similarity in the causal effects (i.e., change in phenotype per allele substitution) across local ancestries within admixed populations; such similarity can be defined as the correlation of ancestral causal genetic effects $r_{\mathrm{admix}} = \mathrm{Cor}[\beta_{\mathrm{afr}}, \beta_{\mathrm{eur}}]$ across African (β_{afr})

and European (β_{eur}) local ancestries. We develop a method that leverages the polygenic architecture of complex traits to model all variants (GWAS-significant and non-significant); this approach is accurate and robust across a wide range of realistic simulated genetic architectures. We also investigate regression-based approaches that use marginal effects of SNPs prioritized in GWAS risk regions. Through simulation studies, we find regression-based methods can yield deflated estimates of similarity (i.e., inflated heterogeneity) especially for highly polygenic traits.

We analyze complex traits in African-European admixed individuals in Population Architecture using Genomics and Epidemiology (PAGE)[16] (24 traits, average *N* = 9K), UK Biobank (UKBB)[43] (26 traits, average *N* = 4K), and All of Us (AoU)[44] (10 traits, average *N* = 20K); there are 38 unique traits in total. We find causal effects are largely consistent across local ancestries within admixed individuals (through meta-analysis across 38 traits, estimated correlation of r_{admix}= 0.95, 95% credible interval [0.93, 0.97]). In addition, we find the heterogeneity in marginal effects exhibited at several trait-locus pairs can be explained by multiple nearby causal variants within a region, consistent with our simulation studies. Our results suggest that the causal effects are largely consistent across local ancestries within African-European admixed individuals, and this motivates future genetic analysis in admixed populations that assume similar effects across ancestries for improved power.

3.2 Results

Overview

We start by describing the statistical model we use to relate genotype to phenotypes in two-way admixed individuals; we focus on two-way African-European admixture because their local ancestries can be accurately inferred (Methods; see Discussion for extension to other admixed

populations). For a given individual, at each SNP s, we denote number of minor alleles from maternal and paternal haplotypes as $x_{s,M}, x_{s,P} \in \{0,1\}$ and local ancestries as $\gamma_{s,M}, \gamma_{s,P} \in \{\text{afr}, \text{eur}\}$. Denoting $\mathbb{I}(\cdot)$ as the indicator function, we define the local ancestry dosage as allele counts from each of ancestries; e.g., $\ell_s = \mathbb{I}(\gamma_{s,M} = \text{afr}) + \mathbb{I}(\gamma_{s,P} = \text{afr})$ for African (similarly for European). For modeling convenience, we use variables that encode the genotypes conditional on local ancestries $g_{s,\text{afr}}, g_{s,\text{eur}}$ as the allele counts specific to each of local ancestries: $g_{s,\text{afr}} \coloneqq x_{s,M}\mathbb{I}(\gamma_{s,M} = \text{afr}) + x_{s,P}\mathbb{I}(\gamma_{s,P} = \text{afr})$ (similarly for $g_{s,\text{eur}}$). The phenotype of an admixed individual is modeled as a function of allelic effect sizes that are allowed to vary across ancestries:

$$y = \sum_{s=1}^{S} \left(g_{s,\text{afr}}\beta_{s,\text{afr}} + g_{s,\text{eur}}\beta_{s,\text{eur}}\right) + \mathbf{c}^{\top}\boldsymbol{\alpha} + \epsilon, \quad (1)$$

where $\beta_{s,\text{afr}}, \beta_{s,\text{eur}}$ are the causal effects at SNP s, S is the total number of causal SNPs in the genome, $\mathbf{c}, \boldsymbol{\alpha}$ are other covariates (e.g., age, sex, genome-wide ancestries) and their effects, and ϵ is the environmental noise. $\beta_{s,\text{afr}}, \beta_{s,\text{eur}}$ are usually referred as *allelic effects*: change in phenotype with each additional allele. This is in contrast with *standardized effects* defined as change in phenotype per standard deviation increase of genotype where genotypes at each SNP s are standardized to have unit variance[45,62]. We refrain from using standardized effects in this work due to complexities arising from different ancestries yielding different ancestry-specific frequencies for the same SNP[45] (Methods).

Our goal is to estimate the similarity in the causal effects across local ancestries in admixed populations (Figure 1); the similarity can be evaluated across all genome-wide causal SNPs that are common across ancestries in a form of cross-ancestry genetic correlation[45,48] (for consistency with previous works we use “genetic correlation” to refer to correlation of genetic effects across

ancestries): $\beta_{s,\mathrm{afr}}, \beta_{s,\mathrm{eur}}$ are modeled as random variables following a bi-variate Gaussian distribution parametrized by σ_g^2, ρ_g, denoting the variance and covariance of the effects:

$$\begin{bmatrix}\beta_{s,\mathrm{afr}}\\ \beta_{s,\mathrm{eur}}\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \tau_s^2 \cdot \begin{bmatrix}\frac{\sigma_g^2}{S} & \frac{\rho_g}{S}\\ \frac{\rho_g}{S} & \frac{\sigma_g^2}{S}\end{bmatrix}\right), s = 1, \ldots, S, \quad (2)$$

where τ_s are variant-specific parameters determined by the genetic architecture assumption (Methods). Under this model, the genome-wide causal effects correlation is defined as $r_{\mathrm{admix}} \coloneqq \frac{\rho_g}{\sigma_g^2}$; $r_{\mathrm{admix}} = 1$ indicates same causal effects across local ancestries, while $r_{\mathrm{admix}} < 1$ indicates differences across ancestries. To estimate r_{admix}, given the genotype and phenotype data for a trait, we calculate the profile likelihood curve of r_{admix}, obtained by maximizing the likelihood of model defined by Equations (1) and (2) with regard to parameters σ_g^2 and environmental variance for each fixed $r_{\mathrm{admix}} \in [0,1]$. We assume $r_{\mathrm{admix}} > 0$ a priori both because that causal effects will unlikely be negatively correlated across ancestries and to reduce r_{admix} search space for reducing computational cost; we have also performed real data analyses to verify this assumption (see below). We obtain the point estimate, credible interval and perform hypothesis testing $H_0: r_{\mathrm{admix}} = 1$ either for each individual trait using the trait-specific profile likelihood curve, or for meta-analysis across multiple traits using the multiplication of the likelihood curves across multiple traits (analogous to inverse variance weighted meta-analysis; Methods).

We organize next sections as follows. First, we show that our proposed approach provides accurate estimation of r_{admix} in extensive simulations. Second, we show r_{admix} is very close to 1 in real data of African-European admixed individuals from PAGE, UKBB and AoU. Third, we replicate our findings using methods that use GWAS summary data (marginal SNP effects at GWAS significant loci). Finally, we investigate pitfalls of methods[22,53,54,63] that use marginal SNP

effects showing inflated heterogeneity; we find that Deming regression is the only approach robust enough to quantify r_{admix} from marginal GWAS effects in admixed individuals.

Polygenic method for r_{admix} is accurate in simulations

We performed simulations to evaluate our proposed polygenic method using real genome-wide genotypes. We simulated phenotypes using genotypes and inferred local ancestries with *N*=17K individuals and S=6.9M SNPs (with MAF > 0.5% in both ancestries in PAGE data set; we omitted population-specific rare SNPs to reduce estimation variance; Methods). Phenotypes were simulated under a range of genetic architectures with a frequency-dependent causal effects distribution[64,65], and varying proportion of causal variants p_{causal}, heritability h_g^2, and true r_{admix} (Methods). We used p_{causal} = 0.1% in our main simulations (to simulate a typical polygenic complex trait[66]). When estimating r_{admix}, we either used all SNPs in the imputed genotypes that were used to simulate phenotypes, or restricted to HapMap3 (HM3) SNPs[67] to simulate scenarios where causal variants are not perfectly typed in the data (Methods).

Our method produced accurate point estimates and well-calibrated credible intervals of r_{admix} across a range of simulation settings (Figure 2a, Supplementary Table 1 and 2). We first evaluated our method in simulations with a realistic range of h_g^2 = 0.1, 0.25, 0.5 and r_{admix} = 0.9, 0.95, 1.0. When using the imputed SNPs for estimation, results were approximately unbiased (average and maximal relative biases across simulation settings were -0.42%, -1.8% respectively). Credible intervals of r_{admix} meta-analyzed across simulations approximately cover true r_{admix}: for the most biased setting (h_g^2 = 0.1, p_{causal} = 0.1%, r_{admix} = 0.95), 95% credible interval = [0.915, 0.948]. When using the HM3 SNPs for estimation, there was a consistent but small downward bias (Figure 2a; average and maximal relative biases were -1.0%, -2.0% respectively). This small

downward bias was due to imperfect tagging that some of the causal SNPs were not included in the HM3 SNPs. Nonetheless, the magnitude of bias using either imputed or HM3 SNPs was small, indicating our method was accurate and robust to imperfect tagging. We next performed simulations to investigate the potential bias in estimating r_{admix} due to omitting population-specific rare variants. We re-applied our methods using SNPs with MAF > 1%, MAF > 5% in both populations (in addition to the default MAF>0.5%) to the same simulated data. We observed downward bias in estimated r_{admix} as more stringent MAF threshold was used and more SNPs were filtered out in estimation procedure. For example, the mode of the estimation was 0.966 when methods were applied with MAF > 5% in simulation of $r_{\mathrm{admix}} = 1.0$ (Figure 2b and Supplementary Table 3). This indicates omitting population-specific rare variants can lead to downward bias (see Discussion). We also investigated the impact of prior assumption of r_{admix}: we applied a revised methodology that allows for $-1 \leq r_{\mathrm{admix}} \leq 1$ and we found that estimated r_{admix} were highly consistent when assuming $0 \leq r_{\mathrm{admix}} \leq 1$ (default method) versus when assuming $-1 \leq r_{\mathrm{admix}} \leq 1$ (Figure 2c).

We performed several secondary analyses. We determined our method remained accurate at other simulated p_{causal} (Supplementary Table 2; p_{causal} ranging from 0.001% to 1%) and broader range of simulated r_{admix} (Supplementary Table 4; r_{admix} ranging from -0.5 to 1). In null simulations (r_{admix} = 1), we determined the false positive rate of hypothesis test $H_0: r_{\mathrm{admix}} = 1$ was properly controlled for most simulation settings, and was only slightly inflated when HM3 SNPs were used, and/or extremely low p_{causal} was simulated. In simulations with r_{admix} < 1, power to detect r_{admix} < 1 increased with increasing h_g^2 and decreasing r_{admix} (Supplementary Table 1 and 2). In addition, we found heritability can be accurately estimated in these simulations

(Supplementary Table 5 and 6; Methods). In summary, our method can be reliably used to estimate r_{admix}.

Causal effects are similar across local ancestries

We applied our polygenic method to estimate r_{admix} within African-European admixed individuals in PAGE[16] (24 traits, average *N*=9296, average fraction of African ancestries=78%), UKBB[43] (26 traits, average *N*=3808, average fraction of African ancestries = 59%), and AoU[44] (10 traits, average *N*= 20496, average fraction of African ancestries = 74%) (see Methods). Meta-analyzing across 38 traits from PAGE, UKBB, AoU (60 study-trait pairs), we observed a high similarity in causal effects across ancestries ($\hat{r}_{\mathrm{admix}}$= 0.95, 95% credible interval= [0.93, 0.97]). Results were highly consistent across data sets despite different ancestry compositions (PAGE: $\hat{r}_{\mathrm{admix}}$= 0.90 [0.85, 0.94], UKBB: $\hat{r}_{\mathrm{admix}}$ = 0.98 [0.91, 1], AoU: $\hat{r}_{\mathrm{admix}}$= 0.97 [0.94, 1]) as well as across traits (Figure 3a, Table 1, Supplementary Table 7). Height was the only trait that had significant $\hat{r}_{\mathrm{admix}} < 1$ (after Bonferroni correction; nominal $p = 4.3 \times 10^{-4} < 0.05/38$; meta-analyzed across three datasets; Table 1) albeit with high estimated $\hat{r}_{\mathrm{admix}} = 0.936\ [0.89, 0.97]$. Estimates of the same traits across datasets were only weakly correlated (Extended Data Figure 1), suggesting similar causal effects by ancestry consistently across traits (true $r_{\mathrm{admix}} \approx 1$ for all traits).

We performed several secondary analyses. Similar to previous simulation studies, we determined prior assumption of r_{admix} had minimal impact to results: estimated r_{admix} of 24 traits in PAGE were highly consistent when assuming $0 \leq r_{\mathrm{admix}} \leq 1$ (default method) versus when assuming $-1 \leq r_{\mathrm{admix}} \leq 1$ (Extended Data Figure 2). Such consistency between the two methods again indicates similar genetic causal effects across local ancestries ($r_{\mathrm{admix}} \approx 1$) and that estimation is robust to choices of statistical prior on r_{admix}. Our results were robust to different assumption of

effects distribution (Extended Data Figure 3 and Supplementary Table 8), consistent with previous work[68]. Results were also robust to the SNP set used in the estimation (Extended Data Figure 3 and Supplementary Table 8), and criterion of the included admixed individuals (Extended Data Figure 4). Additionally, an alternative formulation of method assuming different variance component by ancestry did not outperform our default method assuming same variance component by ancestry (Extended Data Figure 5 and Supplementary Table 9; Supplementary Note).

Next, we contrasted r_{admix} to trans-continental genetic correlations of (1) European vs. African and (2) European vs. East Asian (Figure 3b; Methods). We determine a much higher similarity across local ancestries within admixed populations ($\hat{r}_{\mathrm{admix}}$ = 0.95, 95% credible interval [0.93, 0.97]) as compared to trans-continental correlations of African vs. European within UK Biobank ($\hat{r}_{\mathrm{eur-afr}}$ = 0.50, meta-analysis across 26 traits, 95% confidence interval [0.43, 0.56]) and East Asian (Biobank Japan) vs. European (UK Biobank)[48] ($\hat{r}_{\mathrm{eur-eas}}$ = 0.85, meta-analysis across 31 traits, 95% confidence interval [0.83, 0.87]) (Supplementary Table 10). Overall, our results are consistent with r_{admix} being less susceptible to heterogeneity due to differences in phenotyping/environment in trans-continental comparisons.

We sought to replicate high r_{admix} using regression-based methods that leverage estimated ancestry-specific marginal effects at GWAS loci (Methods). Specifically, we used the following marginal regression equation (restricting Equation (1) to each GWAS SNP s): $y = g_{s,\mathrm{eur}}\beta^{(m)}_{s,\mathrm{eur}} + g_{s,\mathrm{afr}}\beta^{(m)}_{s,\mathrm{afr}} + \mathbf{c}^{\top}\boldsymbol{\alpha} + \epsilon$ (we distinguish marginal effects $\beta^{(m)}$ from causal effects β; Methods). Across 60 study-trait pairs, we detected 217 GWAS significant clumped trait-SNP pairs and we estimated the ancestry-specific marginal effects for each SNP (Figure 3c, Supplementary Table 11). We determined the estimated marginal effects are largely consistent by local ancestry at

these GWAS clumped SNPs via Deming regression slope[69] of 0.82 (SE 0.06) (applied to $\widehat{\beta_{s,\mathrm{eur}}^{(m)}} \sim \widehat{\beta_{s,\mathrm{afr}}^{(m)}}$; Deming regression properly accounts for uncertainty in both dependent and independent variables; Methods). Mean corpuscular hemoglobin (MCH)-associated SNPs at 16p13.3 drove most of the differences by ancestry: Deming regression slope was 0.93 (SE 0.04) on the rest of 193 SNPs after excluding 24 MCH-associated SNPs; MCH-associated SNPs also have the strongest heterogeneity in marginal effects by ancestry (using HET test for effects heterogeneity at each SNP[32]; Supplementary Table 11; Methods). By performing statistical fine-mapping analysis, we found there are multiple conditionally independent association signals at MCH-associated and other loci with heterogeneity by ancestry (Extended Data Figure 6; Supplementary Note). In fact, the MCH-associated loci locate at a region harboring alpha-globin gene cluster (*HBZ-HBM-HBA2-HBA1-HBQ1*) known to contain multiple causal variants[70]. These results suggest that, similar to causal effects, marginal effects at GWAS loci are also largely consistent by local ancestry across multiple traits, with the exception of 16p13.3 loci for MCH in our study, where multiple large-effect causal variants drive some extent of heterogeneity by ancestry in marginal effects.

Pitfalls of using marginal effects to estimate heterogeneity

Next, we focused on thoroughly evaluating methods that use marginal effects at GWAS significant variants to estimate heterogeneity. Marginal effects are frequently used to compare effect sizes across populations or across studies[22,53,54,63] and enjoy popularity for their simplicity and requirement of only GWAS summary statistics (estimated effect sizes and standard errors).

We first note that heterogeneities in marginal effects can be induced due to different LD patterns across ancestries even when the underlying causal effects are identical, especially when multiple

causal variants are nearby in the same LD block (Figure 4). We investigate the extent of heterogeneity by ancestry that can be induced in simulations with identical causal effects across ancestries, due to (1) local ancestry adjustment; (2) unknown causal variants coupled with ancestry-specific LD patterns; (3) highly polygenic genetic architectures with multiple causal SNPs within the same LD block; (4) standard errors in estimated marginal effects across ancestries. Our following simulations were based on real imputed genotypes from African-European individuals in PAGE data (17K individuals, average fraction of African ancestries = 78%).

Regressing out local ancestry can deflate the observed similarity in causal effects across ancestries. We first discuss the use of local ancestry in the heterogeneity estimation, which is a unique and important component to consider when studying admixed populations. We used simulations to investigate the role of local ancestry adjustment using three main approaches: (1) ignoring local ancestry altogether ("w/o"); (2) including local ancestry as covariate in the model ("lanc-included"); (3) regressing out the local ancestry from phenotype followed by heterogeneity estimation on residuals ("lanc-regressed") (Methods). First, in null simulations with identical causal effects (ratio of β_{eur}: $\beta_{\mathrm{afr}} = 1$), we observed that ignoring local ancestry or including local ancestry as covariate yielded well-calibrated HET tests; in contrast, regressing out the local ancestry effect induced inflated HET test statistics (Figure 5 and Supplementary Table 12). Next, in power simulations with varying amount of heterogeneity (defined as ratio of β_{eur}: β_{afr}), including local ancestry in the covariate significantly reduced the power of HET test of up to 50% at high magnitude of heterogeneity (Figure 5 and Supplementary Table 12) (see more details in Supplementary Note). Thus, with respect to local ancestry, we recommend either not using it or

including it as a covariate in the model and not regressing out its effect prior to heterogeneity estimation as that will bias heterogeneity estimation.

Having investigated the role of local ancestry adjustment, we next turn to heterogeneity estimation for GWAS SNPs. We focused on investigating properties of HET test and Deming regression in null simulations with identical causal effects across ancestries ($\beta_{\mathrm{eur}}{:}\,\beta_{\mathrm{afr}} = 1$). Since the true causal variants are usually uncertain, we investigated each method either at the true simulated causal variants or at the LD-clumped variants (Methods).

Uncertainty in which variants are causal can deflate the observed similarity in effects by ancestry. We first performed simulations with single causal variant: we randomly selected 1 SNP as causal in each simulation. Evaluated at the causal SNPs (Methods), we found that HET test and Deming slope were well-calibrated (Figure 6a-c; Extended Data Figure 7; Supplementary Table 13). However, evaluated at the clumped variants, as a more realistic setting (because causal variants need to be inferred), we found HET test became increasingly mis-calibrated with increased h_g^2, while Deming slope remained relatively robust (with an upward but not statistically significant trend with increasing h_g^2). Ordinary least squares (OLS) slope had bias even when evaluated at causal variants because of its ignorance of the standard errors in the estimated effects (Methods and Supplementary Note); such bias became smaller with increased h_g^2.

High polygenicity can deflate the observed similarity in effects by ancestry. Next, we performed simulations where multiple causal variants locate nearby within the same LD block (typical for polygenic complex traits[71,72]; Methods). In this scenario, marginal GWAS effects could tag multiple causal effects thus potentially inflating the observed heterogeneity (Figure 4c). In simulations, we varied the number of causal SNPs from 0.25 to 4.0 per Mb to span most polygenic

architectures. In contrast to simulations with a single causal variant, all three methods (HET test, Deming slope, OLS slope) were biased in the presence of multiple nearby causal variants; the mis-calibration/bias increased with number of causal variants per region. And LD clumping did not alleviate the mis-calibration/bias (Figure 6d-f). Such mis-calibrations occurred irrespective of sample size (Extended Data Figure 8), or simulated heritability h_g^2 (Supplementary Table 14).

In summary, we find that methods for heterogeneity-by-ancestry estimation based on marginal GWAS SNP effects are susceptible to inflated estimates of heterogeneity. HET test is susceptible to false positives when causal variants are unknown. Deming regression was robust in scenarios with low polygenicity, however, was susceptible to inflated estimates of heterogeneity for highly polygenic traits; the inflated estimates can be explained by differential tagging of causal effects across ancestries among causal SNPs. OLS slope had bias because it did not account for uncertainty in estimated effects. We also performed additional simulations with less than identical causal effects ($\beta_{\mathrm{eur}}{:}\beta_{\mathrm{afr}} \neq 1$) and broader range of per-SNP h_g^2 and we determined Deming regression was robust to quantify the heterogeneity level at the marginal effects in simulations of different $\beta_{\mathrm{eur}}{:}\beta_{\mathrm{afr}}$, h_g^2 (Extended Data Figure 9, Supplementary Table 15).

3.3 Discussion

In this work, we developed a polygenic method that model genome-wide causal effects to complex traits of admixed individuals. We determined causal effects are largely similar across local ancestries in analysis of 53K African-European admixed individuals across 38 complex traits in PAGE, UKBB, AoU. In addition to causal effects, we also replicated such consistency-by-ancestry for marginal effects at GWAS loci. We highlighted realistic simulation scenarios where regression-

based methods using marginal effects can report false heterogeneity when causal effects are identical across ancestries.

Our study has several implications for future genetic study of admixed populations, and more broadly of ancestrally diverse individuals. First, reduced accuracy of polygenic score has been observed in African-European admixed populations with increasing proportion of non-European ancestries[56]; our results suggest the causal effects difference has limited contribution to such reduced accuracy. Second, there have been recent work on incorporating local ancestry in statistical modeling of admixed populations, e.g., in association testing[29], polygenic score[56,57], based on the hypothesis that effects may differ across ancestries. Our results indicate the largely consistent causal effects across local ancestries (and also marginal effects at most GWAS loci). The robustness of our results to imperfect tagging also suggests that imperfect tagging induce limited effects heterogeneity across local ancestries, once SNPs are properly modeled in a polygenic model. The small heterogeneity-by-ancestry at causal effects or marginal effects suggest that association tests that do not model heterogeneity-by-ancestry should be preferred in most cases[25,29] for improved statistical power for association. On the other hand, including local ancestry in association models could be useful in correcting for LD induced by admixture[30] and lead to improved causal effect estimation. Full consideration of incorporating local ancestry in statistical models should also take into account the extent of confounding and heterogeneity in the data[73]. Third, our study further motivates studies of ancestrally diverse individuals to identify population-specific risk variants that cannot be investigated due to being rare in European individuals; for example, inclusion of individuals with diverse populations could further disentangle causal from tagging effects thus increasing the power of heterogeneity-by-ancestry estimation. More importantly, larger and robust trans-ancestry studies may allow for the examination of

differential causal effects on a locus-by-locus basis, in addition to the genome-wide approach as presented in this work.

Our results add to the existing literature to further delineate sources of causal effects differences. Previous works have shown moderate causal effects differences across trans-continental populations[45,46,48,63], with part of differences being induced by heterogeneity in the definition of environment/phenotype across continental ancestries. Similarly, a recent work[54] concluded differences between causal effects in European local ancestries within African American admixed individuals and that in European American individuals. Our results showcase that if environments are well controlled (as is the case for genetic variants across local ancestries within admixed populations), causal effects are highly similar across genetic ancestries, agreeing with a recent study finding similar effects across ancestries at level of gene expression in controlled environments[74]. Moreover, our results suggest that local epistatic interaction, if any, does not lead to large causal effects differences across genetic ancestries. By contrasting the high genetic correlation within admixed populations and the low genetic correlation across continental populations, our results support the hypothesis that different environments modify the genetic effects to complex traits (gene-by-environment interaction) across populations.

We note several limitations and future directions of our work. First, we have analyzed SNPs with MAF $\geq$ 0.5% in both ancestries. We excluded population-specific SNPs (with MAF < 0.5% in one of the ancestries) because these SNPs provide little information for estimating r_{admix}, since effects for these SNPs are estimated with large noises. We used simulations to show that omitting these rare variants could lead to downward bias in r_{admix} estimation because of population-specific tagging of shared causal variants (Supplementary Note). However, it remains possible that causal variants themselves are rare and population-specific, and upward bias in the estimation of r_{admix}

may be present. While in this work we focused on estimating r_{admix} for common variants, future work with larger sample sizes is needed to further investigate the impact of population-specific causal SNPs to r_{admix} estimation. Second, we have considered two-way African-European admixed individuals. Several practical considerations remain before applying this method to other admixed populations such as three-way admixture: local ancestries are typically inferred with larger errors[75] and this should be accounted for in statistical modeling (it may be possible to incorporate posterior probabilities in estimated local ancestries to obtain calibrated estimates); additional parameters need to be estimated (e.g., three pairwise correlation parameters across ancestries for three-way admixture populations). We note that our methods can be readily applied to these populations when reliable local ancestry calls can be obtained. Third, our modeling can be extended to estimate correlations in causal effects stratified by functional annotation categories and we leave that as future work. Fourth, our polygenic method requires individual-level genotype and phenotype; if not available, we found Deming regression may be applied to evaluate heterogeneity with caution: in our simulation, Deming regression was the only method robust to most scenarios except for high polygenicity. In our analysis of marginal effects, we found LD clumping can produce cluster of SNPs that were nearby and likely dependent with each other, as a combined result of multiple causal variants within a region and long-range LD in admixed populations. Such dependence may induce bias for methods like Deming regression, highlighting the need for improved methods of identifying conditionally independent SNPs in admixed populations. Fifth, we have meta-analyzed three publicly available studies of PAGE, UKBB, AoU with large cohort of African-European admixed individuals. Such meta-analysis with greatly increased total sample size enabled us to derive the conclusion of the high similarity in causal effects by local ancestry across a broad range of traits. However, our estimates for each individual trait were still associated with large standard errors and can be further improved by analyzing

more individuals. Additional limitations are discussed in the Supplementary Note. Despite these limitations, our study has shown that causal effects to complex traits are highly similar across local ancestries and this knowledge can be used to guide future genetic studies of ancestrally diverse populations.

3.4 Methods

Ethical approval

This research complies with all relevant ethical regulations. Ethics committee/IRB of Population Architecture using Genomics and Epidemiology (PAGE) gave ethical approval for collection of PAGE data. Ethics committee/IRB of UK Biobank gave ethical approval for collection of UK Biobank data (https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/about-us/ethics). Approval to use UK Biobank individual-level in this work was obtained under application 33297 at http://www.ukbiobank.ac.uk. Ethics committee/IRB of All of Us gave ethical approval for collection of All of Us data (https://allofus.nih.gov/about/who-we-are/institutional-review-board-irb-of-all-of-us-research-program). Approval to use All of Us controlled tier data in this work was obtained through application at https://www.researchallofus.org.

Statistical model of phenotype for admixed individuals

For individual $i = 1, \dots, N$ and SNP $s = 1, \dots, S$, we denote $x_{i,s,M}, x_{i,s,P}$ as number of minor alleles at maternal and paternal haplotypes, respectively. We denote corresponding local ancestries as $\gamma_{i,s,M}, \gamma_{i,s,P} \in \{1,2\}$ (we focus on two-way admixture here, e.g., '1' and '2' denote African and European ancestries for African-European admixture). Then we use $g_{i,s,1}, g_{i,s,2}$ to encode allele counts that are specific to each local ancestry:

$$g_{i,s,1} \coloneqq x_{i,s,M}\mathbb{I}(\gamma_{i,s,M}=1)+x_{i,s,P}\mathbb{I}(\gamma_{i,s,P}=1);\quad g_{i,s,2} \coloneqq x_{i,s,M}\mathbb{I}(\gamma_{i,s,M}=2)+x_{i,s,P}\mathbb{I}(\gamma_{i,s,P}=2),$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. Denoting causal allelic effects as $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathbb{R}^S$ for two ancestries, we model the phenotype of each individual y_i as

$$y_i = \mathbf{c}_i^\top \boldsymbol{\alpha} + \sum_{s=1}^{S}\left(g_{i,s,1}\beta_{s,1} + g_{i,s,2}\beta_{s,2}\right) + \epsilon_i, \quad i = 1, \ldots, N$$

where $\mathbf{c}_i \in \mathbb{R}^C, \boldsymbol{\alpha} \in \mathbb{R}^C$ denote C covariates (including all '1' intercepts) and their effects. ϵ_i denotes environmental noise. By further aggregating $g_{i,s,1}, g_{i,s,2}$ into matrices $\mathbf{G}_1 \in \{0,1,2\}^{N\times S}$ and $\mathbf{G}_2 \in \{0,1,2\}^{N\times S}$ for ancestry 1 and 2, and $\mathbf{c}_i$ into $\mathbf{C} \in \mathbb{R}^{N\times C}$, Equation (1) becomes

$$\mathbf{y} = \mathbf{C}\boldsymbol{\alpha} + \mathbf{G}_1\boldsymbol{\beta}_1 + \mathbf{G}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon} \tag{3}$$

We pose the following distribution assumptions $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$ and $\boldsymbol{\epsilon}$

$$\begin{bmatrix}\beta_{s,1}\\ \beta_{s,2}\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \tau_s^2 \cdot \begin{bmatrix}\sigma_g^2/S & \rho_g/S\\ \rho_g/S & \sigma_g^2/S\end{bmatrix}\right), s = 1, \ldots, S \qquad \epsilon_i \sim \mathcal{N}(0, \sigma_e^2), \quad i = 1, \ldots, N \tag{4}$$

where σ_g^2 denotes variance of effects for both populations, ρ_g denotes covariance for similarity of effect sizes by ancestry, and σ_e^2 denotes the variance for environments. τ_s denote SNP-specific parameters (fixed a priori) for effect sizes distribution (see "Specifying τ_s under different heritability models" below). We define correlation of causal genetic effects as $r_{\text{admix}} = \frac{\rho_g}{\sigma_g^2}$. $r_{\text{admix}} = 1$ indicates $\beta_{s,1} = \beta_{s,2}$ for all variants $s = 1, \ldots, S$, i.e., causal effects are the same across ancestries; $r_{\text{admix}} < 1$ indicates differences in causal effects across ancestries.

Calculating and filtering by ancestry-specific allele frequencies. For each SNP s, we calculated MAF as $f_s \coloneqq \frac{\sum_{i=1}^{N}(g_{i,s,1}+g_{i,s,2})}{2N}$. We also calculated ancestry-specific MAF as

$\frac{\sum_{i=1}^{N} g_{i,s,1}}{\sum_{i=1}^{N}[\mathbb{I}(\gamma_{i,s,M}=1)+\mathbb{I}(\gamma_{i,s,P}=1)]}, \frac{\sum_{i=1}^{N} g_{i,s,2}}{\sum_{i=1}^{N}[\mathbb{I}(\gamma_{i,s,M}=2)+\mathbb{I}(\gamma_{i,s,P}=2)]}$ for ancestry 1 and 2. For a SNP s with close-to-zero for either of the ancestry, its effect β_s will be estimated with very large noise. Therefore, we used SNPs with MAF > 0.5% in both ancestries in analyses.

Specifying τ_s under different heritability models. τ_s can model the coupling of SNP effects variance with MAF, local LD or other functional annotations. Commonly used heritability models include GCTA[76], Frequency-dependent[64,65], LDAK[77], and S-LDSC[78] models. While heritability model is important to estimate heritability and functional enrichment of heritability[68,79,80], genetic correlation estimation, the main focus of this study, has shown to be robust to different heritability model[68]. In this work, we mainly used the frequency-dependent model for both simulations and real data analyses (where $\tau_s^2 \propto [f_s(1-f_s)]^{\alpha}$; f_s is the MAF of the SNP s and $\alpha = -0.38$ is estimated in a meta-analysis across 25 UK Biobank complex traits[65]). For real data analysis, we additionally used GCTA model for estimation and found results are robust to heritability models (Extended Data Figure 3).

Alternative choice of genotype normalization by ancestry. We discuss an alternative choice of normalization by ancestry, in which we have two parameters $\tau_{s,1}$ and $\tau_{s,2}$ separately for two ancestries for each SNP. For example, $\tau_{s,1}^2 \propto \frac{1}{f_{s,1}(1-f_{s,1})}, \tau_{s,2}^2 \propto \frac{1}{f_{s,2}(1-f_{s,2})}$ parametrizing effects distribution

$$\begin{bmatrix}\beta_{s,1}\\ \beta_{s,2}\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\tau_{s,1}^2 \cdot \sigma_g^2/S & \tau_{s,1}\tau_{s,2} \cdot \rho_g/S\\ \tau_{s,1}\tau_{s,2} \cdot \rho_g/S & \tau_{s,2}^2 \cdot \sigma_g^2/S\end{bmatrix}\right), s = 1, \dots, S$$

This implies that effects per genotype standard deviation is being modeled (ref.[5] termed this as correlation of allelic impact). While genetic correlation estimation is robust to genotype

standardization (Supplementary Table 8; refs.[45,68]), we recommend modeling allelic effects via same τ_s across ancestries (as used in our default method).

Evaluation of genome-wide genetic effects consistency

We discuss parameter estimation and hypothesis testing in Equations (3) and (4). Marginalizing over random effects $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ in Equation (3), the distribution of $\mathbf{y}$ is

$$\mathbf{y} \sim \mathcal{N}\left(\mathbf{C}\boldsymbol{\alpha}, \sigma_g^2 \frac{\mathbf{G}_1\mathbf{T}\mathbf{G}_1^\top + \mathbf{G}_2\mathbf{T}\mathbf{G}_2^\top}{S} + \rho_g \frac{\mathbf{G}_1\mathbf{T}\mathbf{G}_2^\top + \mathbf{G}_2\mathbf{T}\mathbf{G}_1^\top}{S} + \sigma_e^2\mathbf{I}\right).$$

Where $\mathbf{T}$ is a diagonal matrix with $(\mathbf{T})_{ss} = \tau_s^2$. By denoting $\mathbf{K}_1 = \frac{\mathbf{G}_1\mathbf{T}\mathbf{G}_1^\top + \mathbf{G}_2\mathbf{T}\mathbf{G}_2^\top}{S}$, $\mathbf{K}_2 = \frac{\mathbf{G}_1\mathbf{T}\mathbf{G}_2^\top + \mathbf{G}_2\mathbf{T}\mathbf{G}_1^\top}{S}$, and $\rho_g = \sigma_g^2 \cdot r_{\text{admix}}$, the distribution of $\mathbf{y}$ is simplified as

$$\mathbf{y} \sim \mathcal{N}\left(\mathbf{C}\boldsymbol{\alpha}, \sigma_g^2(\mathbf{K}_1 + r_{\text{admix}}\mathbf{K}_2) + \sigma_e^2\mathbf{I}\right). \quad (5)$$

The maximum likelihood estimates of $(\boldsymbol{\alpha}, \sigma_g^2, r_{\text{admix}}, \sigma_e^2)$ can be found by directly maximizing the corresponding likelihood function $L\left(\boldsymbol{\alpha}, \sigma_g^2, r_{\text{admix}}, \sigma_e^2\right)$. However, the constraint that the correlation parameter r_{admix} should be small than 1 cannot be easily incorporated here. Instead, we use the profile likelihood $L_p(r_{\text{admix}}) := \max_{(\boldsymbol{\alpha}, \sigma_g^2, \sigma_e^2)} L\left(\boldsymbol{\alpha}, \sigma_g^2, r_{\text{admix}}, \sigma_e^2\right)$ and perform grid search of r_{admix} to maximize profile likelihood (similar to ref.[65]): for each candidate r_{admix}, we compute $\mathbf{K}_1 + r_{\text{admix}}\mathbf{K}_2$, and solve $(\boldsymbol{\alpha}, \sigma_g^2, \sigma_e^2)$ for the single variance component model in Equation (5) using GCTA[62] (v1.94.0beta). In practice, we calculate profile likelihood $L_p(r_{\text{admix}})$ for a predefined set of $r_{\text{admix}} = 0.00, 0.05, \ldots, 1.00$ ($r_{\text{admix}} \in [0,1]$ is a reasonable prior assumption here; we alternatively used an extended range of $r_{\text{admix}} = -1, -0.95, \ldots, 0.95, 1.0$ in simulation studies (Supplementary Table 4) and real data analyses (Extended Data Figure 2)). We use natural cubic spline to interpolate pairs

of $\left(r_{\text{admix}}, L_p(r_{\text{admix}})\right)$ to get a likelihood curve of r_{admix}. Then we obtain the estimated $\hat{r}_{\text{admix}}$ using the value that maximize the likelihood curve, and credible interval by combining the likelihood curve with a uniform prior of $r_{\text{admix}} \sim \text{Uniform}[0,1]$ and calculating the highest posterior density interval as credible interval. To perform the meta-analysis across independent estimates, we obtain the joint likelihood by calculating the product of likelihood curves across estimates (or equivalently, the sum of log-likelihood curves), and similarly calculate the estimate and credible interval.

Evaluation of genetic effects consistency at individual variant with marginal effects

Parameter estimation and hypothesis testing. We use a model between individual SNP and phenotype by restricting Equation (1) to the SNP of interest s, as

$$y_i = \mathbf{c}_i^\top \boldsymbol{\alpha} + \left(g_{i,s,1}\beta_{s,1}^{(m)} + g_{i,s,2}\beta_{s,2}^{(m)}\right) + \epsilon_i, \quad i = 1, \dots, N,$$

or in vector form,

$$\mathbf{y} = \mathbf{C}\boldsymbol{\alpha} + \mathbf{g}_{s,1}\beta_{s,1}^{(m)} + \mathbf{g}_{s,2}\beta_{s,2}^{(m)} + \boldsymbol{\epsilon} \tag{6}$$

where $\mathbf{C}, \mathbf{g}_{s,1}, \mathbf{g}_{s,2}, \boldsymbol{\epsilon}$ contain $\mathbf{c}_i, g_{i,s,1}, g_{i,s,2}, \epsilon_i$ for all individuals $i = 1, \dots, N$, respectively. We distinguish marginal effects $\beta_{s,1}^{(m)}, \beta_{s,2}^{(m)}$ in Equation (6) from causal effects $\beta_{s,1}, \beta_{s,2}$ in Equation (1): marginal effects tag effects from nearby causal SNPs with taggability as a function of ancestry-specific correlation between the focal SNP and nearby causal SNPs. Therefore, heterogeneity in marginal effects by local ancestry can be induced even if causal effects are the same (see extensive simulation in Results and more details in Supplementary Note). We estimate $\beta_{s,1}^{(m)}, \beta_{s,2}^{(m)}$ using least squares (jointly for $\beta_{s,1}^{(m)}, \beta_{s,2}^{(m)}$) and perform hypothesis testing of $H_0: \beta_{s,1}^{(m)} = \beta_{s,2}^{(m)}$ with

a likelihood ratio test by comparing Equation (6) to a restricted model where the allelic effects are the same $\beta_s^{(m)} = \beta_{s,1}^{(m)} = \beta_{s,2}^{(m)}$:

$$\mathbf{y} = \mathbf{C}\boldsymbol{\alpha} + \left(\mathbf{g}_{s,1} + \mathbf{g}_{s,2}\right)\beta_s^{(m)} + \boldsymbol{\epsilon} \quad (7)$$

Marginal effects-based methods for estimating heterogeneity. We describe details of marginal effects-based methods to estimate heterogeneity with input from a set of estimated effect sizes $\widehat{\beta_{s,1}^{(m)}}, \widehat{\beta_{s,2}^{(m)}}$ and corresponding estimated standard errors $\mathrm{se}(\widehat{\beta_{s,1}^{(m)}}), \mathrm{se}(\widehat{\beta_{s,2}^{(m)}})$ for a set of SNPs.

Pearson correlation: by calculating the Pearson correlation of $\widehat{\beta_{s,1}^{(m)}}, \widehat{\beta_{s,2}^{(m)}}$ across SNPs. Pearson correlation does not model errors in estimated effects, therefore is expected be smaller than 1 and decreases with increasing error magnitude.

OLS regression slope: by regressing $\widehat{\beta_{s,1}^{(m)}} \sim \widehat{\beta_{s,2}^{(m)}}$ ($\widehat{\beta_{s,1}^{(m)}}$ as dependent variable, $\widehat{\beta_{s,2}^{(m)}}$ as independent variable) or $\widehat{\beta_{s,2}^{(m)}} \sim \widehat{\beta_{s,1}^{(m)}}$. It does not model errors in independent variable. Moreover, it assumes homogeneous errors in dependent variable across SNPs. Therefore, it is susceptible to these error terms and notably results can vary when one exchange the regression orders[81] ($\widehat{\beta_{s,1}^{(m)}} \sim \widehat{\beta_{s,2}^{(m)}}$ vs. $\widehat{\beta_{s,2}^{(m)}} \sim \widehat{\beta_{s,1}^{(m)}}$; e.g., $\widehat{\beta_{s,1}^{(m)}}$ and $\widehat{\beta_{s,2}^{(m)}}$ are associated with different standard errors when being estimated in an admixed population with different ancestry proportion).

Deming regression slope: obtained with Deming regression[69] of $\widehat{\beta_{s,1}^{(m)}}, \widehat{\beta_{s,2}^{(m)}}$ and estimated standard errors $\mathrm{se}(\widehat{\beta_{s,1}^{(m)}}), \mathrm{se}(\widehat{\beta_{s,2}^{(m)}})$. Deming regression models heterogeneous error terms in both independent and dependent variables, therefore is more robust than Pearson correlation and OLS regression. Specifically, given a set of data and estimated standard errors $\left(x_i, y_i, \sigma_{x,i}, \sigma_{y,i}\right), i =$

$1, \ldots, n$ (we use a different set of notations for simplicity), Deming regression optimizes the following objective function to obtain estimated intercept α and slope β:

$$\min_{\substack{\alpha,\beta \\ \delta_1,\ldots,\delta_n \\ \epsilon_i,\ldots,\epsilon_n}} \sum_{i=1}^{n} \left[\frac{\epsilon_i^2}{\sigma_{y,i}^2} + \frac{\delta_i^2}{\sigma_{x,i}^2} \right],$$

$$\text{subject to: } y_i + \epsilon_i = \alpha + \beta(x_i + \delta_i), \quad i = 1, \ldots, n.$$

Standard errors of α, β can be obtained with bootstrapping. Notably, Deming regression slope produce symmetric results with different regression orders (the obtained slope β will be reciprocal to each other). However, Deming regression can still produce biased results when the standard errors $\sigma_{x,i}, \sigma_{y,i}$ are mis-specified[81].

False positive rate of the HET test, as described above in "Parameter estimation and hypothesis testing". It is expected to be well calibrated under the null, because its derivation as a likelihood ratio test. Similar to Deming regression, HET test properly models heterogeneous standard errors.

Genotype data processing

PAGE genotype. We analyzed 17,299 genotyped individuals self-identified as African American in PAGE study[1]. These individuals were from 3 studies: Women's Health Initiative (WHI) (*N*=6,820), Multiethnic Cohort (MEC) (*N*=5,325) and the Icahn School of Medicine at Mount Sinai BioMe biobank in New York City (BioMe) (*N*=5,154). See more details in ref.[16]. The genotypes were imputed to the TOPMed reference panel and we retained well-imputed SNPs with imputation $R^2 > 0.8$ and MAF > 0.5%. We further retained variants with ancestry-specific MAF > 0.5% in both ancestries. This resulted in ~6.9M variants and 17,299 individuals in our analysis.

UK Biobank genotype. We analyzed individuals with African-European admixed ancestries in UK Biobank. We first inferred the proportion of ancestries for each individual in UK Biobank using SCOPE[82] (https://github.com/sriramlab/SCOPE; version Dec. 9th 2021) supervised using 1000 Genomes Phase 3 allele frequencies (AFR, EUR, EAS, SAS). We retained 4,327 African-European admixed individuals with more than 5% of both AFR and EUR ancestries, and with less than 5% of both EAS and SAS ancestries. We retained well-imputed SNPs with imputation R^2 > 0.8 and MAF > 0.5%. We further retained variants with ancestry-specific MAF > 0.5% in both ancestries. This resulted in ~6.6M variants and 4,327 individuals in our analysis.

AoU genotype. We analyzed individuals with African-European admixed ancestries in AoU. We first performed principal component analysis of all 165,208 individuals in AoU microarray data (release v5) joint with 1,000 Genomes Phase 3 reference panel. Then we identified 31,375 individuals with African-European admixed ancestries (with at least both 10% European ancestries and 10% African ancestries, and who was within 2×normalized distance from the line connecting individuals of European ancestries and African ancestries in 1,000 Genomes reference panel; Supplementary Note). For these individuals, we performed quality control using PLINK2[83] (v2.0a3) with `--geno 0.05 --max-alleles 2 --maf 0.001`, and statistical phasing using Eagle2[84] (v2.4.1) with default settings. We retained variants with ancestry-specific MAF > 0.5% in both ancestries. This resulted in ~0.65M variants and 31,375 individuals in our analysis. For AoU, we chose to use microarray data instead of whole genome sequencing data because microarray data of AoU contained more individuals and analyzing microarray data reduced the computational cost.

Local ancestry inference. We performed local ancestry inference using RFMix[85] (https://github.com/slowkoni/rfmix; v2) with default parameters (8 generations since admixture).

We used 99 CEU individuals and 108 YRI individuals from unrelated individuals in 1,000 Genome Project Phase 3[86] as our reference populations, similar to previous works[85,87]. We used HapMap3 SNPs[67] in inference, and then interpolated the inferred local ancestry results to other variants in both PAGE and UK Biobank data sets. The accuracy of RFMix for local ancestry inference has been validated for African-European admixed individuals[29] (e.g., ~98% accuracy for simulations with a realistic demographic model for African American individuals). We performed additional analyses using PAGE African American individuals to assess the robustness of local ancestry inference using an alternative set of reference data. We used all European and African individuals in 1,000 Genomes project (excluding African Caribbean in Barbados (ACB) and African Ancestry in SW USA (ASW) because they were admixed). We determined a high consistency of 98.9% for the inferred local ancestry using reference data of CEU/YRI or all European/African individuals. We used the inferred local ancestry for both simulation study and real data analysis described below.

Simulation study

We describe methods for simulations that corresponds to each section of the Results.

Pitfalls of including local ancestry in estimating heterogeneity. We first describe strategies of including local ancestry in estimating heterogeneity.

For "lanc included", we follow common practices[29–31,88] to use a local ancestry term ℓ_s (defined above) in Equation (1):

$$y = \ell_s \beta_{s,\mathrm{lanc}}^{(m)} + g_{s,1}\beta_{s,1}^{(m)} + g_{s,2}\beta_{s,2}^{(m)} + \mathbf{c}^{\mathsf{T}}\boldsymbol{\alpha} + \epsilon,$$

where $\beta_{s,\mathrm{lanc}}^{(m)}$ denotes the effect of local ancestry.

For "lanc regressed", we use $y = \ell_s \beta_{s,\text{lanc}}^{(m)} + g_{s,1}\beta_{s,1}^{(m)} + g_{s,2}\beta_{s,2}^{(m)} + \epsilon$. We first estimate $\widehat{\beta_{s,\text{lanc}}^{(m)}}$ in the regression of $y \sim \ell_s \beta_{s,\text{lanc}}^{(m)}$, and then estimate $\beta_{s,1}^{(m)}, \beta_{s,2}^{(m)}$ in regression of $\left(y - \ell_s \widehat{\beta_{s,\text{lanc}}^{(m)}}\right) \sim g_{s,1}\beta_{s,1}^{(m)} + g_{s,2}\beta_{s,2}^{(m)}$.

To assess the impact of including local ancestry term when applying HET test, we randomly selected 1,000 SNPs on chromosome 1 from PAGE genotype. We simulated traits with single causal SNP. For each SNP, we simulated quantitative trait with the given single causal SNP with varying $\beta_{\text{eur}}: \beta_{\text{afr}} = 1.0, 1.05, 1.1, 1.15, 1.2$. We scaled $\beta_{\text{eur}}, \beta_{\text{afr}}$ such that the causal SNP explained the given amount of h_g^2. For each SNP, simulations of $\beta_{\text{eur}}, \beta_{\text{afr}}$ and environmental noises were repeated 30 times. We then applied different strategies of including local ancestry to these simulations and obtained *p*-value of HET testing $H_0: \beta_{\text{eur}} = \beta_{\text{afr}}$. We additionally included the top principal component as a covariate throughout. We evaluated the distribution of FPR or power of HET test by sub-sampling *without* replacement: we drew 100 random samples, each sample consisted of 500 SNPs, randomly drawn from the pool of 1,000 SNPs and 30 simulations; such sampling accounts for the randomness from both the environmental noises and SNP MAF. We calculated FPR or power for each sample of 500 SNPs, obtained empirical distributions of FPR or power (100 points each), and then calculated the mean and SE (using empirical standard deviation) from the empirical distribution.

Simulations with single causal variant. We performed simulations with single causal variant to assess the properties of methods based on estimated marginal effects. We randomly selected 100 regions each spanning 20 Mb on chromosome 1 (120K SNPs per region on average, SD 6K). For each region, the causal variant located at the middle of the region; it had same causal effects across local ancestries and was expected to explain a fixed amount of heritability (0.2%, 0.6%,

1.0%); the sign of the causal effect and environmental noises were randomly drawn 100 times. We evaluated 4 metrics at both causal variants and clumped variants; clumped variants were obtained with regular LD clumping (index $p < 5 \times 10^{-8}$, $r^2 = 0.1$, window size = 10 Mb) using PLINK (v1.90b6.24): `--clump --clump-p1 5e-8 --clump-p2 1e-4 --clump-r2 0.1 --clump-kb 10000`. We used a 10Mb clumping window to account for the larger LD window within admixed individuals; other parameters were adopted from ref.[89]. We found that when the simulated h_g^2 was large, LD clumping can result in multiple SNPs because the secondary SNPs can reach $p < 5 \times 10^{-8}$ when we applied a commonly-used $r^2 = 0.1$ threshold. Therefore, for each region, we either retained only the SNP with strongest association (matching the simulation setup of a single simulated causal variant), or retained all the SNPs from clumping results. Similar as above, we evaluated the distribution of 4 metrics by sub-sampling without replacement: we drew 100 random samples, each sample consisted of 500 regions (each region has 1 causal SNP), randomly drawn from the pool of 100 regions and 100 simulations; such sampling accounted for the randomness from both the environmental noises and SNP MAF. We then calculated the mean and SE from the 100 random samples.

Simulation with multiple causal variants. We performed simulations with multiple causal variants. We simulated multiple causal variants randomly distributed on chromosome 1 (515,087 SNPs). We drew n_{causal} = 62, 125, 250, 500, 1000 causal variants to simulate different level of polygenicity, such that on average there were approximately 0.25, 0.5, 1.0, 2.0, 4.0 causal variants per Mb. We fixed the heritability explained by all variants on chromosome 1 as h_g^2 = 2.5%, 5%, 10%, 20%. We performed sub-sampling without replacement to estimate the average and standard errors of 4 metrics (each sample consisted of 1,000 SNPs, randomly drawn from SNPs across 500 simulations). We found that when the simulated h_g^2 was small (h_g^2 = 2.5%, 5%),

because the limited sample size in our data (*n*=17,299) for PAGE data, very few SNPs reach $p < 5 \times 10^{-8}$ in these simulations and consequently standard errors are very large and results cannot be reliably reported. Therefore, we chose to report results only from h_g^2 = 10%, 20% in Supplementary Table 14.

Genome-wide simulation for evaluating our polygenic method. We performed simulations to evaluate our polygenic method in terms of parameter estimation of r_{admix} and hypothesis testing $H_0: r_{\text{admix}} = 1$ using real genome-wide genotypes. We simulated quantitative phenotypes using genotypes and inferred local ancestries from PAGE data set. The phenotypes were simulated under a wide range of genetic architectures varying proportion of causal variants p_{causal}, heritability h_g^2, and true correlation r_{admix}, and a frequency dependent effects distribution for causal variants: in each simulation, we randomly drew p_{causal} proportion of causal variants. Given the set of causal variants, we simulated quantitative phenotypes based on Equations (3) and (4). The environmental noises were then simulated according to the desired heritability h_g^2.

Real data analysis

Phenotype processing. For PAGE, we analyzed 24 heritable traits in PAGE based on ref.[16]. For UK Biobank, we analyzed 26 heritable traits based on heritability and number of individuals with non-missing phenotype values, following ref.[90]. For All of Us, we analyzed 10 heritable traits, including physical measurement and lipid phenotypes, which were straightforward to phenotype and have large sample sizes. Physical measurement phenotypes were extracted from Participant Provided Information in AoU dataset. Lipid phenotypes (including LDL, HDL, TC, TG) were extracted following https://github.com/all-of-us/ukb-cross-analysis-demo-project/tree/main/aou_workbench_siloed_analyses, including extracting most recent measurements per person, and correcting value with statin usage.

These traits included both quantitative and binary traits and it was previously shown that genetic correlation methodology can be directly applied to binary traits[91]. For each trait, we quantile normalized phenotype values. We included age, sex, age*sex, and top 10 in-sample principal components (and "study center" for PAGE) as covariates. We quantile normalized each covariate and used the average of each covariate to imputed missing values in covariates.

Genome-wide genetic correlation estimation. We calculated $\mathbf{K}_1, \mathbf{K}_2$ matrices in Equation (5) using either imputed SNPs and HapMap3 SNPs (for PAGE and UKBB), or microarray SNPs (for AoU). We used either frequency-dependent or GCTA heritability models via specifying τ_s^2. $\mathbf{K}_1, \mathbf{K}_2$ matrices were separately calculated for individuals within PAGE, UKBB, AoU studies. For each given r_{admix}, we used GCTA[62] (v1.94.0beta) to fit a single variance component model with the calculated $\mathbf{K}_1 + r_{\text{admix}}\mathbf{K}_2$ using `gcta64 --reml --reml-no-constrain`. We additionally included the causal signals at Duffy SNP (rs2814778) in 1q23.2 as covariates for analysis of white blood cell count and C-reactive protein because of the known strong admixture peak[92,93]. Specifically, we used the local ancestries of SNP closest to Duffy SNP in our data as proxies for Duffy SNP (Duffy SNP itself is not typed or imputed in our data). The local ancestries are valid proxies of Duffy SNP because Duffy SNP is known to be highly differentiated across ancestries (alternate allele frequency is 0.006 vs. 0.964 in ref.[86]) and therefore local ancestries are highly correlated with the Duffy SNP. We excluded closely related individuals in the analysis (< third-degree relatives; using ref.[94] with `plink2 --king-cutoff 0.0884`). We note that our meta-analysis credible interval across traits can be anti-conservative (i.e., the actual coverage probability is less than the nominal coverage probability) because we did not account for the genetic correlation across traits.

Individual trait-SNP analysis. We evaluated effects consistency at individual SNPs that were significantly associated with each trait. First, we performed GWAS and LD clumping with the same parameters described above. Even though LD clumping was performed using stringent parameters, we found cluster of clumped SNPs that were likely dependent with each other as a combined result of multiple causal variants within a region the long-range LD in admixed populations (Supplementary Table 11; Discussion). For each clumped trait-SNP pair, we estimated ancestry-specific effects and standard errors.

Statistical fine-mapping analysis. We performed fine-mapping analysis to each trait-SNP pair with significant heterogeneity by ancestry using SuSiE[95] (v0.12) (for PAGE and UKBB, for which we used genotype data with high SNP density). For each trait-SNP, we included all imputed SNPs in 3Mb window. We ran SuSiE with individual-level genotype and phenotype (covariates were regressed out of genotype and phenotype), using default settings with maximum number of 10 non-zero effects. We obtained posterior inclusion probability and credible sets.

Statistics and reproducibility

We analyzed three publicly available datasets of PAGE, UK Biobank and All of Us and sample sizes were determined from these studies. We did not use randomization or blinding. We focused on analyzing individuals with admixed African-European ancestries and individuals with other genetic ancestries were not included in analyses of this work. We replicate our findings across these three independent datasets.

3.5 Figures

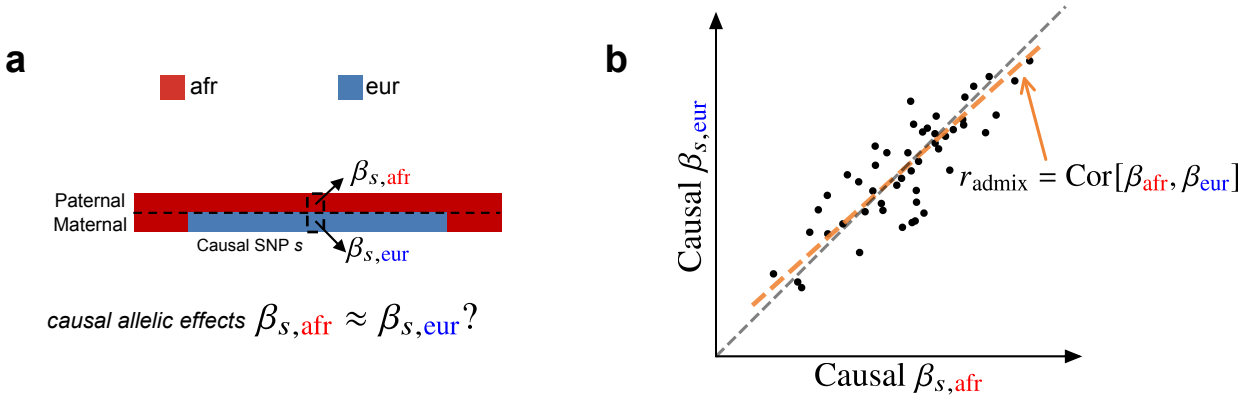

Figure 3.1 Concepts of estimating similarity in the causal effects across local ancestries.

(a) For a given trait, with phased genotype (paternal haplotype at the top and maternal haplotype at the bottom) and inferred local ancestry (denoted by color), we investigate whether $\beta_{s,\mathrm{afr}} \approx \beta_{s,\mathrm{eur}}$ across each causal SNP s. **(b)** We focus on estimating the genome-wide correlation of genetic effects across ancestries $r_{\mathrm{admix}} = \mathrm{Cor}[\beta_{\mathrm{afr}}, \beta_{\mathrm{eur}}]$, which is the regression slope (orange line) of ancestry-specific causal effects. For reference, the grey dashed line corresponds $\beta_{\mathrm{afr}} = \beta_{\mathrm{eur}}$.

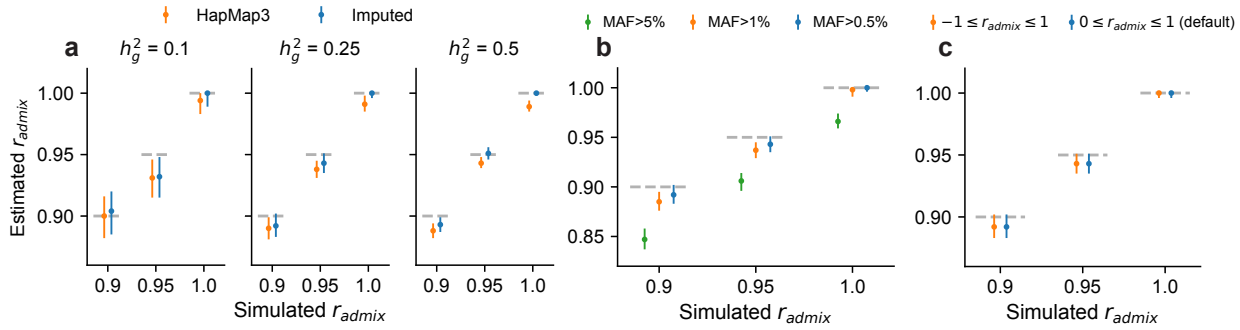

Figure 3.2 Results of genetic correlation r_{admix} estimation in genome-wide simulations.

Simulations were based on 17K PAGE individuals and 6.9M genome-wide imputed variants with MAF > 0.5% in both ancestries. We fixed the proportion of causal variants p_{causal} as 0.1% and varied genetic correlation r_{admix} = 0.90, 0.95, 1.0. **(a)** Impact of using HapMap3 or imputed variants in estimation. We varied simulated genome-wide heritability h_g^2= 0.1, 0.25, 0.5. **(b)** Impact of selecting common variants at different MAF thresholds in estimation. h_g^2 was fixed to 0.25 and imputed variants at different MAF thresholds were used in estimation. **(c)** Impact of prior assumption in estimation. h_g^2 was fixed to 0.25 and imputed variants were used in estimation. For each simulated genetic architecture, we plot the mode and 95% credible interval based on the meta-analysis across 100 simulations (Methods). Numerical results are reported in Supplementary Table 1-4 (including results for other $p_{\text{causal}}, r_{\text{admix}}$).

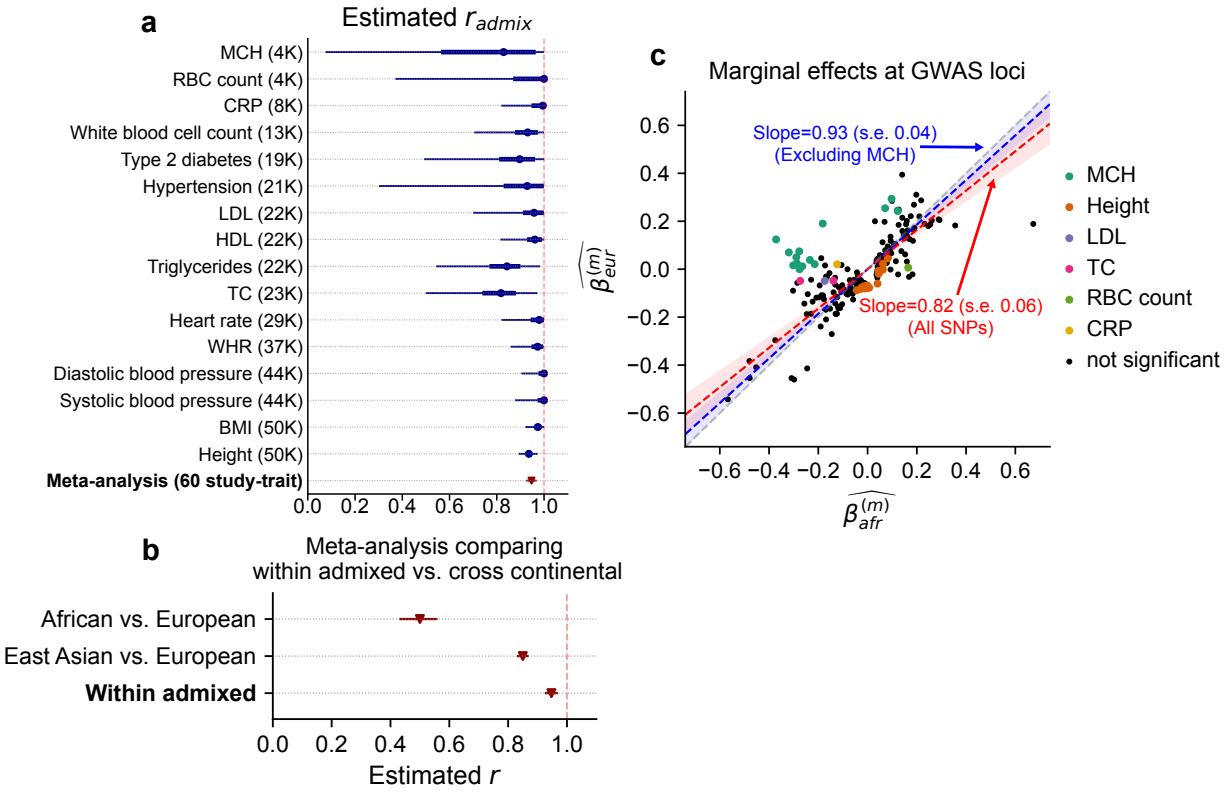

Figure 3.3 Similarity of causal effects and marginal effects across local ancestries meta-analyzed across PAGE, UKBB, AoU.

(a) We plot the trait-specific estimated r_{admix} for 16 traits. For each trait, dots denote the estimation modes; bold lines and thin lines denote 50% / 95% highest density credible intervals, respectively. Traits are ordered according to total number of individuals included in the estimation (shown in parentheses). These traits are selected to be displayed either because they have the largest total sample sizes, or because the associated SNPs of these traits exhibit heterogeneity in marginal effects (see the panel on the right). We also display the meta-analysis results across 60 study-trait pairs (38 unique traits). Numerical results are provided in Table 1. **(b)** Comparison of r_{admix} (n=38 traits) to meta-analysis results from trans-continental genetic correlation of African vs. European (n=26 traits) and East Asian vs. European (n=31 traits). Point estimates and 95% confidence intervals are denoted using triangles and lines. **(c)** We plot the ancestry-specific marginal effects for 217 GWAS significant clumped trait-SNP pairs across 60 study-trait pairs. Trait-SNP pairs with significant heterogeneity in marginal effects by ancestry (p_{HET}< 0.05/217 via HET test) are denoted in color (non-significant trait-SNP pairs denoted as black dots; some black dots with large differences across ancestries were not significant because of the large standard errors in estimated effects). Numerical results are reported in Supplementary Table 11. Point estimates and 95% confidence intervals for Deming regression slopes of $\widehat{\beta_{s,\mathrm{eur}}^{(m)}} \sim \widehat{\beta_{s,\mathrm{afr}}^{(m)}}$ are provided either for all 217 SNPs (red), or for 193 SNPs after excluding 24 MCH-associated SNPs (blue). MCH, mean corpuscular hemoglobin. RBC, red blood cell. CRP, C-reactive protein. LDL, low density lipoprotein cholesterol. HDL, high density lipoprotein cholesterol. TC, total cholesterol. BMI, body mass index. WHR, waist to hip ratio.

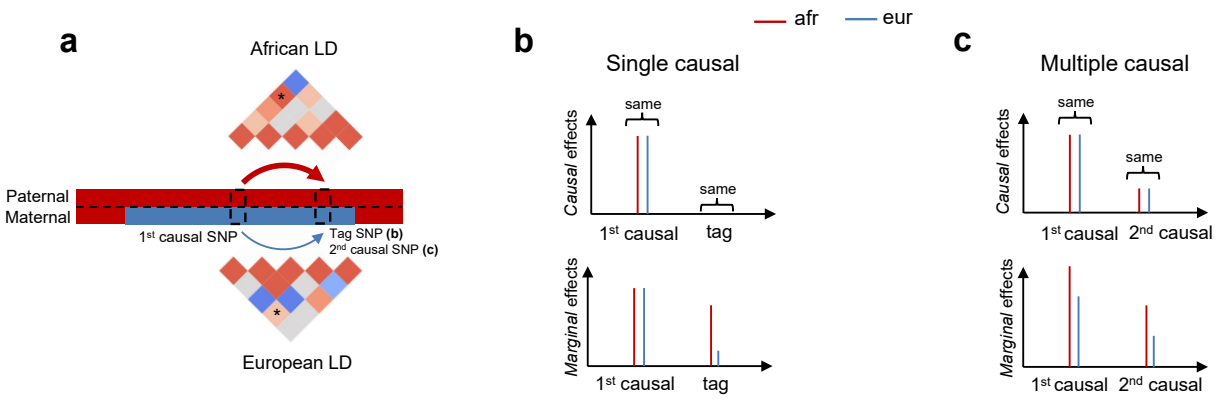

Figure 3.4 Induced heterogeneities in marginal effects across local ancestries.

(a) Illustrations that different LD patterns across local ancestries can induce differential tagging between a causal SNP and a tag SNP in **(b)** or another causal SNP in **(c)**. LD strengths between the two SNPs are indicated both in the thickness of arrows and in the color shades of '*' elements in LD matrices. **(b)** Example of single causal SNP with no heterogeneity. Causal effects are the same across local ancestries, and the estimated marginal effects at causal SNP will be also very similar with sufficient sample size. However, because of differential tagging across local ancestries, the estimated marginal effects evaluated at the tag SNP are different. **(c)** Example of multiple causal SNPs with no heterogeneity. Causal effects for both SNPs are the same across local ancestries. In this example, the correlation between the 2 causal variants is higher for genotypes in African local ancestries than those in European local ancestries. Therefore, African ancestry-specific genotypes tag more effects, creating different ancestry-specific marginal effects at each causal SNP.

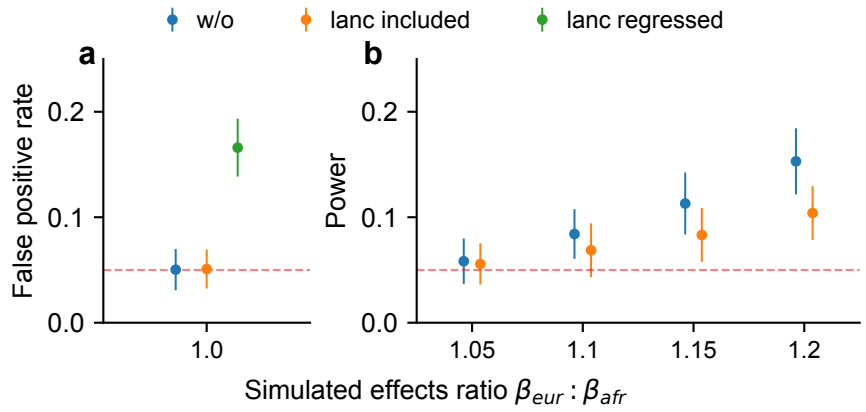

Figure 3.5 Pitfalls of including local ancestry in estimating heterogeneity.

In each simulation, we selected a single causal variant and simulated quantitative phenotypes where these causal variants explain heritability h_g^2 = 0.6%; we also varied ratios of effects across ancestries β_{eur}: β_{afr}. **(a)** False positive rate in null simulation β_{eur}: β_{afr}= 1.0. **(b)** Power to detect $\beta_{\text{eur}} \neq \beta_{\text{afr}}$ in power simulations with β_{eur}: β_{afr}> 1. We did not include "lanc regressed" because it is not well-calibrated in null simulations. We plot the mean and 95% confidence intervals, calculated via 100 random sub-samplings with each sample consisting of 500 SNPs (Methods). Numerical results are reported in Supplementary Table 12.

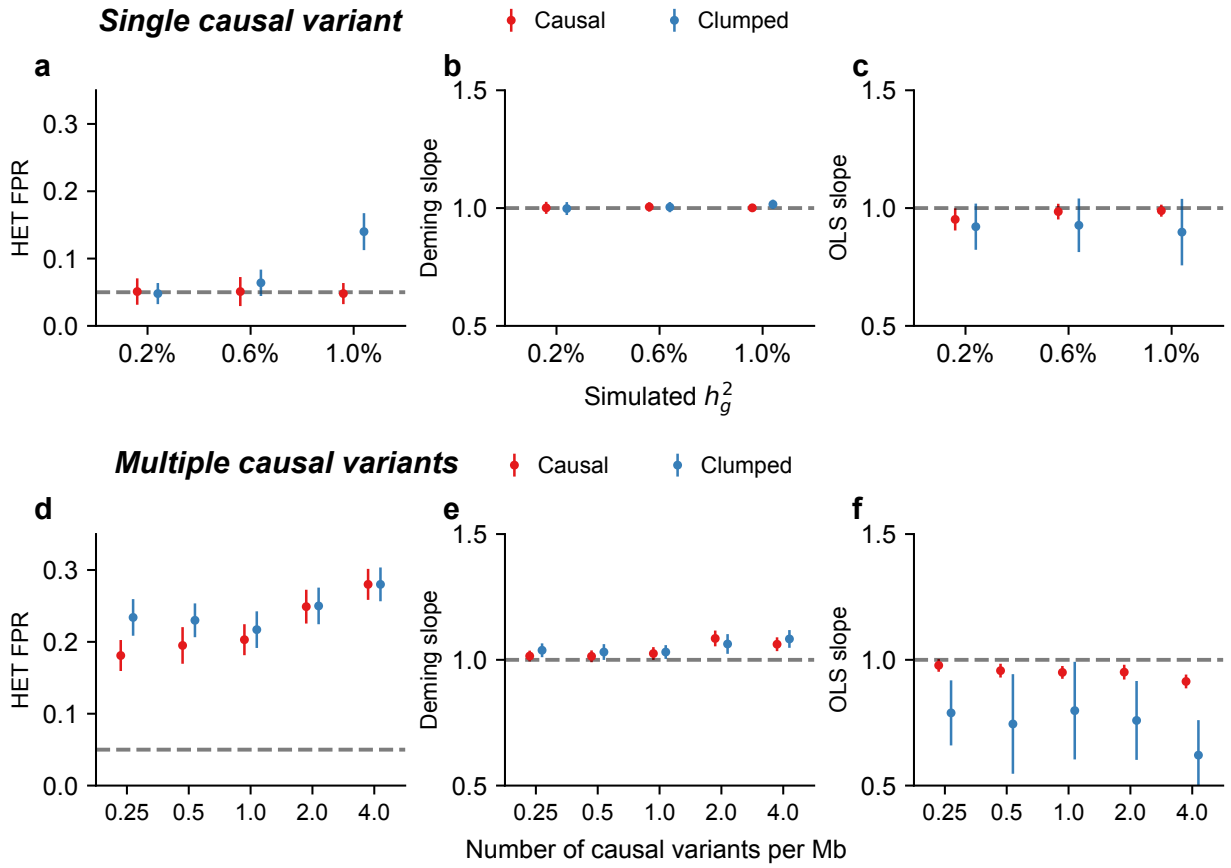

Figure 3.6 Mis-calibration of HET test / Deming regression / OLS regression in simulations with r_{admix} = 1.

(a-c) Simulations with single causal variant. Each causal variant had the same causal effects across local ancestries and each causal variant explained a fixed amount of heritability (0.2%, 0.6%, 1.0%). **(a)** False positive rate (FPR) of HET test. **(b)** Deming regression slope of $\widehat{\beta_{\text{eur}}^{(m)}} \sim \widehat{\beta_{\text{afr}}^{(m)}}$. **(c)** OLS regression slope of $\widehat{\beta_{\text{eur}}^{(m)}} \sim \widehat{\beta_{\text{afr}}^{(m)}}$. Numerical results are reported in Supplementary Table 13. **(d-f)** Simulation with multiple causal variants. We simulated different level of polygenicity, such that on average there were approximately 0.25, 0.5, 1.0, 2.0, 4.0 causal variants per Mb. Causal variants had same causal effects across local ancestries. The heritability

explained by all causal variants was fixed at $h_g^2 = 10\%$. **(d)** FPR of HET test. **(e)** Deming regression slope of $\widehat{\beta_{\text{eur}}^{(m)}} \sim \widehat{\beta_{\text{afr}}^{(m)}}$. **(f)** OLS regression slope of $\widehat{\beta_{\text{eur}}^{(m)}} \sim \widehat{\beta_{\text{afr}}^{(m)}}$. 95% confidence intervals were based on 100 random sub-samplings with each sample consists of 1,000 SNPs (Methods).

3.6 Tables

Trait	*N*	$\hat{r}_{\mathbf{admix}}$ mode	95% credible interval(s)	*p*-value	$\widehat{h_g^2}$
BMD	1668	0.000	[0.00, 0.78]	0.012	0.34±0.16
Neuroticism	3044	1.000	[0.36, 1.00]	1	0.36±0.11
Education years	3324	0.000	[0.00, 0.94]	0.4	0.055±0.075
MCHC	3650	0.228	[0.00, 0.87]	0.061	0.21±0.092
Type 1 diabetes	3767	0.381	[0.00, 0.95]	0.77	-0.033±0.016
HLR count	3852	1.000	[0.07, 1.00]	1	0.12±0.086
RBC distribution width	3925	1.000	[0.27, 1.00]	1	0.28±0.087
Lymphocyte count	3935	1.000	[0.00, 0.60] [0.66, 1.00]	1	0.13±0.086
Monocyte count	3935	0.972	[0.26, 1.00]	0.82	0.3±0.087
MCH	3948	0.829	[0.07, 1.00]	0.36	0.2±0.076
RBC count	3948	1.000	[0.37, 1.00]	1	0.31±0.09
Hypothyroidism	4063	1.000	[0.05, 1.00]	1	0.046±0.07

PR interval	4071	0.844	[0.08, 1.00]	0.36	0.22±0.084
QRS interval	4078	1.000	[0.07, 1.00]	1	0.12±0.082
Asthma	4079	1.000	[0.15, 1.00]	1	0.21±0.087
Ever smoked	4083	0.764	[0.04, 0.98]	0.31	0.17±0.082
QT interval	4089	0.920	[0.07, 1.00]	0.69	0.16±0.083
HbA1c	5353	0.954	[0.08, 1.00]	0.77	0.19±0.078
Cigarettes per day	6995	0.999	[0.08, 1.00]	1	0.097±0.047
Fasting insulin	7753	1.000	[0.21, 1.00]	1	0.13±0.044
eGFR	7978	0.805	[0.16, 1.00]	0.09	0.19±0.046
C-reactive protein	8321	0.995	[0.82, 1.00]	0.94	0.28±0.046
Fasting glucose	9646	0.695	[0.00, 0.93]	0.27	0.064±0.035
Coffee consumption	11587	0.982	[0.10, 1.00]	0.9	0.074±0.03
Platelet count	12545	0.783	[0.20, 0.98]	0.025	0.19±0.038
White blood cell count	12755	0.931	[0.70, 1.00]	0.26	0.23±0.036

Type 2 diabetes	18630	0.897	[0.49, 1.00]	0.23	0.12±0.024
Hypertension	20744	0.929	[0.30, 1.00]	0.45	0.08±0.027
LDL	21979	0.958	[0.70, 1.00]	0.55	0.14±0.046
HDL	22039	0.961	[0.82, 1.00]	0.46	0.22±0.057
Triglycerides	22494	0.843	[0.54, 0.98]	0.012	0.18±0.027
Total cholesterol	22555	0.818	[0.50, 0.97]	0.007	0.18±0.039
Heart rate	28764	0.980	[0.82, 1.00]	0.74	0.099±0.015
WHR	36756	0.973	[0.86, 1.00]	0.55	0.12±0.015
Diastolic blood pressure	43787	1.000	[0.90, 1.00]	1	0.077±0.024
Systolic blood pressure	43788	1.000	[0.88, 1.00]	1	0.071±0.013
BMI	49521	0.974	[0.92, 1.00]	0.33	0.22±0.02
Height	49605	0.936	[0.89, 0.97]	0.00043	0.4±0.014
Meta analysis		0.947	[0.93, 0.97]	8.7×10^{-7}	

Table 3.1 Genome-wide genetic correlation across 38 complex traits for African-European admixed individuals in PAGE, UKBB, AoU.

For each trait, we report number of individuals, posterior mode and 95% credible interval(s) for estimated r_{admix}, nominal one-sided *p*-value for rejecting the null hypothesis of $H_0: r_{\mathrm{admix}} = 1$ (unadjusted for multiple testing; Methods), and estimated heritability and standard error. Meta analysis results performed across 38 traits are shown in the last row. Traits are ordered according to number of individuals. For each trait, we perform meta-analysis across studies if the trait is in multiple studies (Methods). Lymphocyte count has two credible intervals because of the non-concave profile likelihood curve, as a result of small sample size. BMD, bone mineral density. HLR, high light scattering reticulocytes. MCHC, mean corpuscular hemoglobin concentration.

4 Calibrated prediction intervals for polygenic scores across diverse contexts

4.1 Introduction

Accurate prediction of complex diseases/traits integrating genetic and non-genetic factors is essential for a wide range of fields from agriculture to personalized genomic medicine. The genetic contribution to traits is typically predicted using polygenic scores (PGS) that summarize the joint contribution of many genetic factors[96–99]. A critical barrier in PGS use is their *context-specific accuracy* – their performance (and/or bias) varies across genetic ancestry[18,56,100–102], age, sex, socioeconomic status and other factors[103–105]. This prevents equitable use of PGS across individuals of all contexts[18,99,106].

PGS use large-scale genome-wide association studies (GWAS) to estimate linear prediction models of traits based on genetic variants; these models are then used for new data that often has different context characteristics from the GWAS training data (e.g., different distributions of genetic ancestry, social determinants of health)[96,97,107]. Even when testing data is similar to training data, genetic effects themselves can vary by contexts (e.g., due to genotype-environment interaction, across age[108], sex[109], genetic ancestry[45,48,54,110]) thus leading to differential PGS performance (as traditional PGS do not model such interactions). Furthermore, when genetic effects are unknown, allele frequency, linkage disequilibrium and differential tagging of true latent genetic factors can also lead to context-specific accuracy of PGS-based predictions[50,103,108].

To account for PGS accuracy variability, we propose to incorporate context-specificity using *trait prediction intervals* that vary across contexts. Trait prediction intervals denote the range

containing true trait values with pre-specified confidence (e.g., 90%). They provide a natural approach to model variability in PGS accuracy – narrower prediction intervals correspond to contexts where PGS attains higher accuracy – that can then be utilized in PGS-based predictions[103,111,112]. As an example, consider the case of two individuals with the same PGS-based predictions for low-density lipoprotein cholesterol (LDL) of 180 mg/dL. If the two individuals have different contexts (e.g., sex) that are known to impact PGS accuracy (e.g., R^2=0.1 in men vs. 0.2 in women), their prediction intervals will also vary (e.g., 180 ± 40 mg/dL vs. 180 ± 10 mg/dL). In this example, the second individual is more likely to meet a decision criterion of LDL>160 mg/dL for clinical intervention.

To achieve calibration across all contexts, we propose a statistical framework (*CalPred*) that jointly models the effects of all contexts on PGS accuracy leveraging calibration data. The key assumption is that new target individuals for whom PGS-based predictions will be employed have similar distribution of contexts as calibration data. This is motivated by precision health efforts that created EHR-linked biobanks of patients from the same medical system in which PGS-based predictions will be implemented in the future[113–116]; in this context the assumption is that the biobank is representative of future patients entering the same medical system.

First, we analyze data from two large-scale biobanks (UK Biobank[43] and All of Us[24]) to find pervasive impact of context on PGS accuracy across a wide range of traits. All considered traits (N=72) have at least one context impacting their accuracy[103,105]. Socio-economic contexts have similar magnitudes of impact on PGS accuracy as genetic ancestry; for example, PGS accuracy varies by up to ~50% for individuals across “education years” context averaged across all considered traits in All of Us. Socio-economic contexts have greater impact on PGS accuracy in All of Us, a more diverse dataset, as compared to UK Biobank. Our results can be used to identify

important contexts to account for when implementing PGS-based prediction in diverse populations.

Second, we use simulations and real data analysis to find that CalPred provides calibrated predictions across individuals of diverse contexts. For quantitative traits, CalPred jointly models the impact of genetic ancestry, age and sex and other social determinants of health. In LDL prediction, prediction intervals need adjustment by up to ~40% across contexts to achieve calibration. Context-specificity of PGS prediction varies across traits and the studied population; for example, prediction intervals for education years need adjustment by 94% in All of Us versus 10% in UK Biobank, reflecting the more diverse distribution of education years and other social determinants of health in All of Us. For disease traits, incorporating context information is critical for calibrated predicted probability. In All of Us, PGS-based type 2 diabetes (T2D) predictions ignoring “annual household income” are mis-calibrated across income groups, while incorporating income in the model leads to calibrated predictions. Overall, our approaches provide a path forward to developing and applying PGS for human trait predictions across diverse contexts.

4.2 Results

Overview

We incorporate context-specific accuracy in PGS-based predictions using prediction intervals varying across contexts to maintain calibration: the true phenotype is contained within the prediction interval at a pre-specified probability (e.g., 90%; Fig. 1a). Naturally, as accuracy varies by context, the interval width needs to vary adaptively to maintain calibration (Fig. 1b). For illustrative purpose we distinguish among three types of prediction intervals (Fig. 1c). First, standard errors of PGS weights can be used to estimate prediction intervals that do not vary

across contexts and/or individuals; these types of intervals are calibrated only when target perfectly matches training which is hard to achieve in practice. Second, prediction intervals can be estimated empirically using a calibration dataset while ignoring context[96,117–121]; these types of intervals are robust to mismatches between training and testing, but are mis-calibrated in particular contexts due to the variability of PGS accuracy. Third, prediction intervals that vary across contexts can be estimated using a calibration dataset by empirically quantifying the impact of each context on prediction accuracy; context-specific prediction intervals are adaptive and robust across contexts albeit at the expense of a more complex statistical model and larger calibration data that spans all contexts.

We distinguish three categories of datasets when calibrating predictions. *Training data*, used to perform GWAS and PGS weights estimation, often involves meta-analysis of multiple datasets where additional context adjustment is impractical due to data access limitations or unmeasured context variables. *Calibration data* is used to calibrate PGS with respect to trait-relevant contexts. For example, EHR-linked biobanks within medical systems are generated in part to calibrate PGS-based predictions before any clinical implementation. *Testing data* refers to new individuals for which the calibrated prediction models will be employed (e.g., patients within medical systems not currently involved in EHR-linked biobanks). Motivated by clinical implementation of PGS-based predictions in medical systems where EHR-linked biobanks already exist, here we focus on leveraging calibration data to estimate context-specific prediction intervals. In this scenario it is natural to use existing EHR-linked biobanks as approximation for future patients within the same medical system; therefore, our approach assumes that calibration and testing data are from similar populations. For example, UCLA ATLAS biobank[113] contains data of UCLA Health patients that can be used to calibrate PGS-based predictors for future visits of UCLA patients. Our approach does not require training and calibration data to match in contexts (Discussion).

Context-specific prediction intervals are implemented with two components: (1) *context-specific mean* $\hat{y}_i = \mathbb{E}[y_i|\mathbf{c}_i]$ as a function of context $\mathbf{c}_i$ for each individual *i*; we also include PGS-Context interaction terms (PGSxC) to model varying PGS slope across contexts in this work; (2) *context-specific variance* $\mathbb{E}[(y_i - \hat{y}_i)^2|\mathbf{c}_i] = \exp\left(\mathbf{c}_i^\top \boldsymbol{\beta}_\sigma\right)$, where $\mathbf{c}_i$ denotes contexts including age, sex, socioeconomic factors and top principal components as major axes of genetic ancestry and $\boldsymbol{\beta}_\sigma$ quantifies the unique impact of each context on variation of the prediction interval accounting for other contexts (Methods). Denoting prediction standard deviation (SD) as $\hat{\sigma}_i = \sqrt{\exp\left(\mathbf{c}_i^\top \hat{\boldsymbol{\beta}}_\sigma\right)}$, 90% prediction intervals can be derived as $(\hat{y}_i - 1.645 \times \hat{\sigma}_i, \hat{y}_i + 1.645 \times \hat{\sigma}_i)$. Our approach builds upon existing models for heteroscedasticity in probabilistic forecasting[122–126]. Existing works incorporate variable residual variances across different subsets of data (i.e., contexts in our case) in addition to modeling prediction mean in standard regression analysis. Within genetics literature, such models have been used to detect genotypes associated with phenotype variability (vQTL)[127–129]. Here, we build on such methods towards modeling PGS variable accuracy across contexts.

Widespread context-specific PGS accuracy across populations

Although PGS accuracy has been shown to vary across selected traits and contexts[18,103–105], its pervasiveness remains unclear. We analyzed two large-scale biobanks in the UK and US (UK Biobank and All of Us) comprising >600K individuals spanning a wide range of contexts. We trained PGS for 72 traits in individuals previous annotated as "White British"[43] (WB) from UK Biobank and evaluated these PGSs in independent testing data from UK Biobank and All of Us. We focused on 11 contexts that span genetic ancestry, sex, age, and socio-economic factors such as educational attainment (Methods). We used *relative* ΔR^2 to quantify the impact of context to PGS accuracy defined as $\frac{R^2_{\text{top quintile}} - R^2_{\text{bottom quintile}}}{R^2_{\text{all}}}$, where $R^2_{[\text{subset}]}$ denotes R^2 between PGS

and residual phenotype computed in a given range of the context variable (top/bottom quintile for continuous contexts; binary subgroups for binary contexts). We found widespread context-specific PGS accuracies across all traits and contexts studied (Fig. 2, Supplementary Fig. 1 and 2, Supplementary Table 1 and 2; Methods).

Context-specific accuracy in UK Biobank

All 72 traits had at least one context impacting their accuracy in UK Biobank data; 264 (out of 792) PGS-context pairs had significant variable accuracy ($p < 0.05 / (72 \times 11)$; Methods). Overall, genetic ancestry had the most widespread impact on PGS accuracy: 70 of 72 traits had significant differences in PGS accuracy, with an average relative ΔR^2 of -46% between top and bottom PC1 quintiles (Supplementary Fig. 3). Socioeconomic contexts also significantly impacted PGS accuracy; PGS accuracy significantly differed for 62 traits, with an average relative ΔR^2 of -23% between top and bottom deprivation index quintiles. The direction of context's impact depended on the trait being studied. For example, age significantly impacted 20 traits; rather than consistently increasing or decreasing accuracy, an older age led to increased accuracy for 14 traits (e.g., high-density lipoprotein cholesterol and white blood cell count in Fig. 2; HDL and WBC) and to decreased accuracy for 6 traits (e.g., LDL).

The widespread context-specificity remained when testing data was matched to training data by genetic ancestry (Fig. 2). 21 (out of 72) PGSs had at least one context significantly impacting their prediction accuracy; 42 PGS-context pairs had significant variable accuracy ($p < 0.05 / (72 \times 11)$). We replicated previously reported variable PGS accuracy in WB individuals for diastolic blood pressure, body mass index, education years across contexts of sex, age and deprivation index[103].

As an example, LDL was significantly impacted by six contexts in WB individuals, with age having the strongest impact (relative ΔR^2 was more than 100% between top and bottom age quintiles).

Next, we studied the unique impact of each context on variable PGS accuracy within CalPred model jointly accounting for all contexts (Methods, Fig. 2cd). Context contribution to variable accuracy conditional on all other contexts was quantified with β_σ, where larger absolute β_σ indicated more substantial variation in accuracy along a context variable (Methods). In general, effects of contexts to traits were largely independent. For example, both PC1 and deprivation index significantly impacted PGS accuracy for a range of traits in the joint model, indicating both had a unique contribution to variable PGS accuracy. We also found examples showing otherwise: the impact of "wear glasses" context on LDL accuracy can be explained by its correlation with age (Extended Data Fig. 1), while other contexts independently contributed to variable LDL accuracy. These results indicated the importance of jointly considering all measured contexts to correctly assess the unique contribution of each context. We found that contexts including sex, age, income, and deprivation index had comparable impact on accuracy as genetic ancestry (Fig. 2ef). The distribution of estimated effects of β_σ suggested predominantly higher prediction accuracy for individuals with higher income and lower deprivation indices; this can be partly explained by different context distribution PGS training data: WB individuals had higher income and lower deprivation indices compared to the rest of the UK Biobank[130] (Extended Data Fig. 2). We noted two context-trait pairs with large differences between single-context and combined-context analysis results even within UK Biobank white British individuals (sex-BMI and sex-WHR). This is because single-context analysis uses population-level R^2 focusing on the predictive power of only PGS while combined-context analysis assesses the impact of context on phenotypical residual variance (Supplementary Note).

Context-specific accuracy in All of Us

We next turned to All of Us, a diverse biobank across the US comprising more than 245K participants (Supplementary Fig. 3 and Extended Data Fig. 3). Due to challenges in phenotype matching across biobanks, we restricted the analysis to 12 PGS and 11 contexts matching the UK Biobank analyses (Methods). All PGS had at least one context impacting their accuracy (Fig. 3, Supplementary Table 3 and 4). 89 PGS-context pairs were significant when considering all individuals, and 61 PGS-context pairs were significant when restricting to individuals with self-reported race/ethnicity (SIRE) as "White" ("White SIRE") ($p < 0.05 / (12 \times 11)$; Methods). Prediction of cholesterol and LDL were similarly impacted by a broad range of contexts. Prediction of education years was impacted by contexts including age, BMI, employment, income, both when considering all individuals and considering "White SIRE" sample, consistent with that socioeconomic contexts influence PGS of socio-behavioral traits such as education[103,131,132].

Interestingly, socioeconomic contexts had greater impact on context-specificity in All of Us as compared to UK Biobank. For example, years of education context significantly impacted 9 out of 11 traits with average relative ΔR^2=50%, as compared to 2 out of 71 traits with average relative ΔR^2=0.2% in UK Biobank (averaging across traits other than education years itself). This may be explained by larger variation of education years in the US and/or education being more correlated with social determinants of health in the US compared to the UK. When restricting analysis to subset of individuals with more homogenous genetic ancestry, the impact of contexts of education years and income level was attenuated but remained significant; this is consistent with variable PGS accuracy across socioeconomic contexts being partially accounted for through their correlation with genetic ancestry (Extended Data Fig. 4).

For completeness we also evaluated PGSs for height[133] and LDL[134] derived from multi-ancestry meta-analyses from PGS Catalog[135] (Fig. 3). We found that multi-ancestry PGSs did not alleviate widespread context-specific accuracy. Higher income, education years, better employment, or lower BMI predominately led to higher prediction accuracy across traits (Fig. 3ef). We formally compared and determined an overall consistency for fitted β_σ coefficients across populations and biobanks (Supplementary Fig. 4). We determined that variable R^2 across contexts was not solely driven by differences of phenotype variance in context strata: context-specific R^2 can result from differences in either phenotypic variance or PGS predictiveness, and the extent attributed to either component varied by each context-trait pair (Supplementary Fig. 5). We further verified that context-specificity patterns remained significant when context variables themselves were regressed out from the initial GWAS for PGS training (Supplementary Fig. 6).

CalPred is calibrated across contexts in simulations

Having shown that context-specificity of PGS accuracy is pervasive across traits and biobanks, we next turned to CalPred to estimate context-specific prediction intervals accounting for context- and trait-specific variable accuracy (Methods). We performed simulations to evaluate calibration of CalPred in the presence of gene-by-context interactions[109,136]. For quantitative traits, we simulated individuals in two contexts with different heritability and an imperfect genetic correlation (the first context is used to train PGS; Methods; Fig. 4a). Due to genetic heterogeneity, PGS weights derived in the first context were not portable to the second context, producing a biased phenotype-PGS regression slope and prediction intervals with deflated coverage. With CalPred, prediction mean was calibrated via PGSxC terms; prediction interval lengths were adjusted to reflect different prediction precision across two contexts. For disease traits, we simulated individuals in two contexts under a liability threshold model with different disease prevalence and

an imperfect genetic correlation (Fig. 4b; Methods). We first predicted disease probability with a logistic regression model for all individuals in both contexts, using PGS weights derived from the first context. As expected, this model ignoring context information was mis-calibrated overall in each context. By incorporating PGS, PGSxC interaction and context variables, we determined disease risk predictions were then calibrated within and across contexts. We also simulated other scenarios of gene-context interactions for both quantitative and disease traits, and verified that our framework produced calibrated predictions (Extended Data. Fig. 5 and 6).

We next evaluated CalPred in simulations where prediction accuracy varies across contexts similar to real data[18,100,103] (Fig. 5; Methods). We assessed calibration of prediction intervals both at the overall level and within each context subgroup (Methods). First, generic prediction intervals without context-specific adjustment had severe over-/under-coverage within each context subgroup stratified by PC1, age, or sex. As expected, bias of coverage tracked closely with accuracy across contexts. Second, CalPred context-specific prediction intervals were calibrated across contexts, by incorporating context-specific prediction accuracy in the interval estimation. We also performed simulations to find CalPred performance depended on calibration sample size N_{cal}>500 for accurate model fitting, and it is important to select an appropriate set of contexts in calibration (Extended Data Fig. 7). Parameter estimation of $\boldsymbol{\beta}_{\sigma}$ was accurate with correctly-specified model and remained robust in model mis-specification scenarios (Supplementary Fig. 7). Overall, simulation results demonstrated that CalPred produces well-calibrated prediction intervals when contexts are measured and present in the data, and highlighted the importance of comprehensive profiling of relevant context information.

CalPred yields calibrated context-specific predictions

We applied CalPred to produce context-specific prediction intervals for a wide range of quantitative traits across UK Biobank and All of Us. We first performed several analyses in All of Us to investigate best practices to model quantitative traits. We examined effects of PGS, context variables, and PGSxC for trait prediction and found that PGS had the largest contribution in explaining trait variation (cross-trait average standardized effects with magnitudes of 0.23 compared to 0.22 of sex and 0.14 of BMI, the second and third largest contributors). PGSxC had significant contributions but with smaller effects than those from context variable themselves (Extended Data Fig. 8). Notably, inclusion of PGS substantially increased inter-individual variation in prediction SD, suggesting that PGS is an important source of variation in prediction accuracy across individuals (Extended Data Fig. 9). PGSxC and VbyC components had additive contribution in improving model fitting, indicating they modeled independent aspects of traits (Supplementary Fig. 8-10).

We next focus on LDL, an important risk factor of cardiovascular disease[134]. Calibration by context is particularly important because LDL prediction accuracy was impacted by many contexts, with largest impact from age (Fig. 2 and 3). We modeled prediction mean using PGS together with age, sex, and genetic ancestry, and modeled context-specific prediction intervals using the set of contexts in Fig. 2 and 3 (Methods). LDL prediction accuracy decreased with age (R^2=18% in youngest quintile vs. R^2=11% in oldest quintile; Fig. 6a). Generic prediction intervals were mis-calibrated with coverage of 93% and 86% for youngest and oldest quintiles instead of the nominal level of 90%. In contrast, context-specific prediction intervals had the expected 90% coverage across all considered contexts. This resulted from varying prediction interval length by context, with a wider interval compensating for lower prediction accuracy. For example, as CalPred estimated a positive impact of age to prediction uncertainty (β_σ=0.15; $p<10^{-30}$), individuals in

youngest/oldest age quintiles had average prediction standard deviation (SD) of 27.4 vs. 34.3 mg/dL (25% difference; Supplementary Fig. 11; Methods). These findings were replicated in All of Us and in other traits (Supplementary Fig. 12 and 13), where R^2 varied across contexts and context-specific prediction intervals achieved well-calibration across contexts providing per-individual accuracy metrics (Supplementary Fig. 14). Next, we sought to examine the joint contribution of all considered contexts to variable prediction SD (instead of separately considering age, PC1 or sex; Fig. 6b). Context-specific accuracy was more pronounced by ranking individuals by prediction SD accounting for impact of all contexts (prediction SD ranged approximately from 20 mg/dL to 45mg/dL; Fig. 6b): we detected a 44% difference comparing individuals in bottom and top deciles of prediction SD (25.2 mg/dL vs. 36.5 mg/dL; Fig. 6c; Supplementary Fig. 15 and 16). This implied that individuals in top prediction SD decile (characterized by contexts of male, increased PC1 and age; see Fig. 2c) needs to have prediction interval widths increased by 44% compared to those in bottom decile.

Extending analysis accounting for all contexts to all traits in UK Biobank and All of Us, we determined a widespread large variation of context-specific prediction intervals across traits (Fig. 7 and Supplementary Fig. 17). Average differences between top and bottom prediction SD deciles across traits were 30% and 47%, respectively for UK Biobank and All of Us. The trait with the highest prediction SD difference was the average mean spherical equivalent (avMSE), a measure of refractive error, that was impacted the most by "wear glasses" context. Individuals who wore glasses had a much higher PGS-phenotype R^2 than those who did not, likely due to the reduced variation in avMSE phenotypes among individuals who did not wear glasses. Comparing across the two datasets, BMI, LDL, and cholesterol were more heavily influenced by context than average, while diastolic blood pressure and HDL were less impacted, suggesting trait-specific susceptibility to context-specific accuracy. There were cases where context-specificity of the

same trait was drastically different across datasets. For example, prediction SD differences for predicting education years was 94% in All of Us versus 10% in UK Biobank. This disparity likely reflected the more diverse distribution of education years and other social determinants of health in the US population sampled in All of Us (Fig. 2 and 3). Such differences between datasets also highlight that context-specificity can be population-specific and the need to consider characteristics of different populations in calibration.

We next investigated disease risk prediction for four well-powered heritable diseases: type 2 diabetes (T2D)[137], coronary artery disease (CAD)[138], prostate cancer (PrCA)[139], breast cancer (BrCA)[140] (Extended Data Fig. 10). We first considered a baseline model using logistic regression to predict disease probability with PGS, age, sex, BMI, top 10 PCs as predictors (Methods). We evaluated calibration of predicted disease risk – whether predicted probability aligns with observed disease rate. While baseline model predictions were calibrated at an aggregate level, they were mis-calibrated within specific contexts (Fig. 8a). For example, among individuals with a predicted T2D risk of approximately 30% (25%-35%, N=4,662), the observed proportion with T2D was 30.9% (S.E.=0.7%). However, this proportion varied significantly with individual's "annual household income": 32.7% (S.E.=2.0%) in the lowest income bracket (N=562) had T2D, compared to only 18.1% (S.E.=2.3%) in the highest income bracket (N=271); T2D risk was consistently under-estimated for individuals of lower income and over-estimated for individuals with higher income. The discrepancy suggests that a baseline model ignoring disease-relevant contexts produces severely mis-calibrated probability estimates. We then used a logistic model to incorporate contexts including "annual household income" together with their interaction with PGS, to find that predicted disease risk was calibrated at the overall level and also within each income group; modeling variance by context for disease liability achieved similar calibration (Fig. 8b, Supplementary Fig. 18 and 19); we discussed reasons explaining their similar performances

(Methods). Overall, our results emphasize the importance of incorporating contexts into probability risk calibration to achieve calibrated predictions across all considered contexts.

4.3 Discussion

Our work adds to the literature of PGS-based prediction as follows. We show that context-specific accuracy of PGS is highly pervasive across traits and biobanks with socioeconomic contexts often having larger impact than genetic ancestry[18,19,103,105,112]. We introduce CalPred to estimate context-specific prediction intervals. Compared with other PGS calibration approaches, CalPred incorporates context information leveraging a calibration dataset (Supplementary Note). For quantitative traits, CalPred provides a framework to quantify individualized context-specific generalizability/portability of a given PGS to be leveraged in downstream analyses. Prediction intervals can be interpreted as a reference range accounting for each individual's contexts (including age, sex, and genetic variation via PGS). They provide individual-level uncertainty metrics improving PGS applications. For example, they can be used to identify individuals having PGS-based predictions with exceedingly high uncertainty and inform cases when it is not appropriate to report polygenic scoring results because of the high instability. For disease traits, we found models which overlooked context information resulted in mis-calibrated disease probability predictions in the presence of gene-context interactions. Such miscalibrations are problematic if they lead to over-/under-diagnosis for individuals across socioeconomic context groups. To address this, proper incorporation of context variables and PGSxC interactions in PGS-based predictions led to calibrated predictions across contexts.

We note several limitations of our work. First, we motivated our approach for clinical implementation using continuous biomarkers; we focused on LDL as an example continuous lab

value with clinical application. Other biomarkers to consider could be prostate-specific antigen (PSA) currently employed for patient stratification for biopsies and prostate cancer diagnosis. Recent work has highlighted incorporating genetically predicted PSA levels improves clinical utility by reducing unnecessary biopsies and improving detection of aggressive form of prostate cancer[14]. Therefore, lab values form a useful system for prediction method development that may have clinical implications; actual clinical utility requires thorough implementation considering clinical decision processes, and we leave that as future work. Second, CalPred requires calibration data that matches in distribution with the target data, including both distribution of contexts and their impact to traits. Otherwise, there may be bias in target samples underrepresented in calibration data. Meanwhile, we note that PGS weights do not need to be derived from the same population as the target population. For example, in this work, PGS weights trained in white British individuals from UK Biobank achieved expected calibration in All of Us. The calibration process can incorporate differences across PGS training and target population into context-specific prediction intervals. Third, comprehensive profiling of context information is fundamental in applying calibration and interpreting results. In our simulation studies, missing contexts prevent proper calibration of PGS. In our T2D analysis, “annual household income”, instead of a causal context, may be a proxy to other contexts such as diet and physical exercise that are more directly relevant to T2D. We advocate standardized and comprehensive profiling of contexts across biobanks to quantify the role of contexts to PGS accuracy. Relatedly, GWAS data collection needs to prioritize diversity not only in genetic ancestry, but also across social-economic contexts. Fourth, we found limited improvement of model fitting when including PGS in VbyC and therefore did not include PGS because model interpretation is more straightforward when prediction variance is solely a function of contexts. However, it is technically valid to include PGS in VbyC as genetic contexts may modify prediction precisions. Fifth, context-specific accuracy

can arise due to biological genetic effects differences across contexts such as GxAge and GxSex, or because of statistical differences of MAF/LD patterns contributing to a substantial proportion of PGS performance differences across genetic ancestry. Disentangling various aspects driving context-specific accuracy is an ongoing research direction[103,109,136]. Sixth, this work has primarily focused on the impact of PGS on the variability of prediction intervals across contexts. However, it is important to note that variable accuracy of other predictors and variable phenotypic variance also contribute to our findings. The results presented here regarding variable prediction accuracy should be attributed to the collective impact of all predictors, rather than solely to PGS. While we have determined the substantial contribution of PGS to variable accuracy, further quantifying the relative contributions of each predictor is an important future direction.

4.4 Methods

Ethical approval

This research complies with all relevant ethical regulations. Ethics committee/IRB of UKBB gave ethical approval for collection of UKBB data (https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/about-us/ethics). Approval to use UKBB individual level in this work was obtained under application 33297 at http://www.ukbiobank.ac.uk. Ethics committee/IRB of AoU gave ethical approval for collection of AoU data (https://allofus.nih.gov/about/who-we-are/institutional-review-board-irb-of-all-of-us-research-program). Approval to use AoU controlled tier data in this work was obtained through application at https://www.researchallofus.org.

Constructing calibrated and context-specific prediction intervals

We first provide an overview of CalPred framework. CalPred takes as input pre-trained PGS weights, genotype, phenotype and contexts to train a calibration model producing calibrated and context-specific prediction intervals for target individuals. We consider a calibration dataset with N_{cal} individuals. For each individual i=1, …, N_{cal}, we have genotype vector $\mathbf{g}_i \in \{0,1,2\}^M$ with M SNPs, and phenotype y_i. Using pre-trained PGS weights for a given trait $\boldsymbol{\beta}_g \in \mathbb{R}^M$, we calculate PGS in calibration data with $\mathbf{g}_i^\top \boldsymbol{\beta}_g$. PGS and other contexts including age, sex, genetic ancestry and socioeconomic factors, compose each individual i's contexts $\mathbf{c}_i$ (all '1' intercepts are also included). Phenotypes are modeled as

$$y_i = \mathcal{N}\big(\mu(\mathbf{c}_i), \sigma^2(\mathbf{c}_i)\big), i = 1, \dots, N_{\text{cal}}$$

$$\mu(\mathbf{c}_i) = \mathbf{c}_i^\top \boldsymbol{\beta}_\mu, \sigma^2(\mathbf{c}_i) = \exp\big(\mathbf{c}_i^\top \boldsymbol{\beta}_\sigma\big).$$

There are two main components:

- $\mu(\mathbf{c}_i) = \mathbf{c}_i^\top \boldsymbol{\beta}_\mu$ models the baseline prediction mean using predictors of PGS, contexts, as well as PGSxContext interaction terms (PGSxC).
- $\sigma^2(\mathbf{c}_i) = \exp\big(\mathbf{c}_i^\top \boldsymbol{\beta}_\sigma\big)$ models context-specific variance of y around prediction mean. Differential prediction accuracy across contexts lead to variable variance around prediction mean across contexts. The use of $\exp(\cdot)$ is to ensure that the variance term >= 0. PGSxC terms are not included for ease of interpretation.

We estimate $\boldsymbol{\beta}_\mu, \boldsymbol{\beta}_\sigma$ leveraging calibration data using restricted maximum likelihood for linear model with heteroskedasticity[56] (`statmod` v1.5.0[57]). Individual-specific predictive distribution $\mathcal{N}\big(\hat{\mu}(\mathbf{c}_i) = \mathbf{c}_i^\top \widehat{\boldsymbol{\beta}}_\mu, \widehat{\sigma^2}(\mathbf{c}_i) = \exp(\mathbf{c}_i^\top \widehat{\boldsymbol{\beta}}_\sigma)\big)$ can be generated for any target individual $\mathbf{c}_i$ using the

fitted $\widehat{\boldsymbol{\beta}}_{\mu}, \widehat{\boldsymbol{\beta}}_{\sigma}$. The corresponding α-level prediction interval (e.g., α=90% for 90% prediction interval) is

$\left[\hat{\mu}(\mathbf{c}_i) - \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\hat{\sigma}(\mathbf{c}_i), \hat{\mu}(\mathbf{c}_i) + \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\hat{\sigma}(\mathbf{c}_i)\right]$, where Φ^{-1} is the inverse cumulative distribution function of a standard normal distribution (e.g., $\Phi^{-1}\left(1-\frac{\alpha}{2}\right) = 1.645$ for 90% prediction interval). With moderate sample size for calibration data (e.g., N_{cal}>500 as validated in our simulation studies), such models can be estimated with high precision.

Quantile normalization for non-normal phenotype distribution. In the above, we have assumed that prediction intervals can be modeled as a Gaussian distribution, which may not be valid for every phenotype. For robust implementation in real data, we apply a transformation function $Q(\cdot)$ to y with ranked-based inverse normal transformation such that $Q(y)$ follows a normal distribution; $Q(y)$ can then be modeled using methods described above. Fitted prediction intervals can then be transformed back into the original y space using $Q^{-1}(y)$.

Model for disease trait within the liability threshold model. CalPred model can be extended for disease traits. We first use CalPred to model continuous disease liability $y^{\text{liab}} = \mathcal{N}\big(\mu(\mathbf{c}), \sigma^2(\mathbf{c})\big)$, and then integrate out scenarios where disease liability is above the threshold $P(y=1) = \Phi\big(y^{\text{liab}} > 0\big) = \Phi\left(\frac{\mu(\mathbf{c})}{\sigma(\mathbf{c})}\right)$, where $\Phi(\cdot)$ is a link function used in logistic or probit regression. Intuitively, this maps the continuous liability into disease risk while accounting for liability uncertainty. Disease trait probability can be alternatively modeled using a logistic regression model $P(y=1) = \Phi\big(\mathbf{c}^{\top}\boldsymbol{\beta}_{\text{logistic}}\big)$. In real data analysis (Fig. 8 and Supplementary Fig. 19), we did not observe substantial improvement of CalPred model over logistic linear model. To explain this, comparing logistic regression model $P(y=1) = \Phi\big(\mathbf{c}^{\top}\boldsymbol{\beta}_{\text{logistic}}\big)$ with CalPred model

$P(y=1) = \Phi\left(\frac{\mu(\mathbf{c})}{\sigma(\mathbf{c})}\right)$, we note that $\mathbf{c}^\top \boldsymbol{\beta}_{\text{logistic}}$ can be seen as first-order terms in Taylor expansion approximating $\frac{\mu(\mathbf{c})}{\sigma(\mathbf{c})}$. Therefore, our observation is explained by the fact that linear logistic regression model is a good approximation of CalPred disease model.

Quantifying context-specific R^2 of PGS

We quantify context-specific prediction accuracy (R^2) of PGS, that is, to what extent PGS have variable prediction accuracy across contexts (including age, sex, genetic ancestry, socioeconomic factors that can influence traits[58]). Identification of contexts contributing to variable prediction accuracy is important in constructing calibration model. For each pair of context and trait in a population, we calculated prediction accuracy R^2 between PGS $\hat{y}_i$ and covariate-regressed phenotypes y_i (phenotypes for each trait were regressed out of age, sex, age*sex and top 10 PCs; this adjustment is to better separate the contribution of PGS) across each subgroup of individuals defined by contexts. We summarized results using relative differences of R^2 across context groups to baseline R^2 calculated across all evaluated individuals (differences between two classes for binary contexts; differences between top and bottom quintiles for continuous contexts). We calculated Spearman's R^2 between point predictions and covariate-regressed phenotypes $R^2(\hat{y}, y)$ within each context subgroup. We also calculated the baseline Spearman's R^2 denoted as R^2_{all} across all individuals regardless of contexts. We summarized the results for each pair of trait and context using the "relative ΔR^2" defined as $\frac{R^2_{\text{group1}} - R^2_{\text{group2}}}{R^2_{\text{all}}}$. We assessed statistical significance of ΔR^2 across context subgroups by testing the null hypothesis $H_0: \Delta R^2 = 0$ using 1,000 bootstrap samples of ΔR^2 (in each bootstrap sample, the whole dataset was resampled with replacement and ΔR^2 were then re-evaluated). Statistical significance was assessed using two-sided *p*-values comparing the observed ΔR^2 to the bootstrap samples of ΔR^2.

Relationship between CalPred model and R^2. Population-level metrics such as R^2 can be derived from the model as a function of $\boldsymbol{\beta}_\sigma$ and distribution of $\mathbf{c}_i$. Suppose $y = \hat{y} + e, e \sim \mathcal{N}(0, \exp(\mathbf{c}^\top \boldsymbol{\beta}_\sigma))$, where $y, \hat{y}, e$ denote phenotypes, point predictions and residual noises. We have

$$R^2(y, \hat{y}) = R^2(\hat{y} + e, \hat{y}) = \frac{\mathrm{Var}[\hat{y}]}{\mathrm{Var}[\hat{y}] + \mathrm{Var}[e]}$$

Holding $\mathrm{Var}[\hat{y}]$ as fixed, $R^2(y, \hat{y})$ is a function of $\mathrm{Var}[e]$, which is determined by the distribution of $\mathbf{c}$ and values of $\boldsymbol{\beta}_\sigma$. This indicates a correspondence between $\boldsymbol{\beta}_\sigma$ and $R^2(y, \hat{y})$. Therefore, estimated $\boldsymbol{\beta}_\sigma$ can also be used as a metric to quantify context-specific accuracy (as used in Fig. 2 and 3). While relative ΔR^2 is easier to interpret, it assesses the marginal contribution of each context separately and require discretization of continuous contexts. Meanwhile, β_σ in CalPred model jointly account for all contexts in parametric regression, and therefore can quantify the unique distribution of each context to variable accuracy. On the other hand, even with constant prediction interval length (constant $\mathrm{Var}[e]$), variable R^2 can result from variable $\mathrm{Var}[\hat{y}]$ across context groups. While CalPred focuses on modeling $\mathrm{Var}[e]$ as a function of contexts to represent variable R^2, $\mathrm{Var}[\hat{y}]$ can change across contexts. For example, $\mathrm{Var}[\hat{y}]$ can vary with contexts if $\hat{y} = \mathrm{PGS} \times \beta_{\mathrm{slope}}$ and the slope β varies as a function of context. Such variable slope term can be modeled with variable slope terms in prediction mean $\hat{y}$ (Supplementary Note).

Real data analysis

We analyzed a diverse set of contexts and traits in UK Biobank and All of Us (1) to quantify the extent of context-specific prediction accuracy; (2) to evaluate context-specific prediction intervals via CalPred for quantitative traits; (3) to evaluate probability prediction for disease traits.

Polygenic score weights. Polygenic scores were trained on 370K individuals in UK Biobank that were assigned to "white British" cluster and 1.1M HapMap3[59] SNPs. For each trait, we performed GWAS using PLINK2 (v2.0a3) `plink2 --glm` with age, sex and top 16 PCs as covariates. We estimated PGS weights using `snp_ldpred2_auto` in LDpred2[60] (bigsnpr v1.8.1) with GWAS summary statistics and in-sample LD matrix. These PGS weights were applied to target individuals in both UK Biobank and All of Us to obtain individual-level PGS. To train PGS weights for All of Us individuals, we overlapped 1.2M SNPs in All of Us quality-controlled microarray data to 12M SNPs in UK Biobank imputed data to obtain a set of 0.8M SNPs present in both datasets. Then we trained and applied PGS weights using these shared SNPs in UK Biobank to All of Us individuals. This procedure improves PGS accuracy in All of Us by ensuring all SNPs with non-zero weights to present in the data.

UK Biobank dataset. We analyzed 490K genotyped individuals (including both training and target individuals). We used 1.1M HapMap3[59] SNPs in all analyses. All UK Biobank individuals are clustered into sub-continental ancestry clusters based on top 16 pre-computed PCs (data-field 22009 in ref.[28] as in ref.[6]). This procedure assigned 410K individuals into "white British" cluster. A random subset of 370K "white British" individuals to perform GWAS and estimate PGS weights (see above); we trained PGS weights starting with individual-level data to avoid overlap of sample between training and target data. For evaluation, we used the rest of 120K individuals with genotypes, phenotypes and contexts (including individuals from both ~40K "White British" individuals and ~80K other individuals). We focused on analyzing 72 traits with R^2>0.05 in 40K WB target individuals and/or biological importance). We followed https://github.com/privefl/UKBB-PGS/blob/main/code/prepare-pheno-fields.R and ref.[6] to preprocess trait values (e.g., log-transformation and clipping of extreme values). For each trait, we quantile normalized phenotype

values; when performing calibration, phenotype quantiles were calculated based on calibration data and then used to normalize target data. We analyzed 11 contexts representing a broad set of socioeconomic and genetic ancestry contexts, including binary contexts (sex, ever smoked, wear glasses, drinking alcohol) and continuous contexts (top two PCs, age, BMI, income, deprivation index, and education years). We note that income and education years have been processed into 5 quintiles in the original data of UK Biobank.

All of Us dataset. We analyzed 245K genotyped individuals with diverse genetic ancestry contexts (short read whole genome sequencing data in release v7). We retained 1.2M SNPs from microarray data after quality control using PLINK2 (v2.0a3) with `plink2 --geno 0.05 --chr 1-22 --max-alleles 2 --rm-dup exclude-all --maf 0.001`. We used microarray data because it contains more individuals and can be analyzed with low computational cost. All individuals with microarray data were used in the evaluation. We analyzed 10 heritable traits, including height, BMI, WHR, diastolic blood pressure, systolic blood pressure, education years, LDL, cholesterol, HDL, triglyceride; they are straightforward to phenotype and have large sample sizes. Physical measurement phenotypes were extracted from Participant Provided Information. Lipid phenotypes (including LDL, HDL, TC, TG) were extracted following https://github.com/all-of-us/ukb-cross-analysis-demo-project/tree/main/aou_workbench_siloed_analyses, including procedures of extracting most recent measurements per person, and correcting for statin usage. For each trait, we quantile normalized phenotype values; when performing calibration, phenotype quantiles were calculated based on calibration data and were then used to normalize target data. We included age, sex, age*sex, and top 10 in-sample principal components as covariates in the model. We also quantile normalized each covariate and used the average of each covariate to impute missing values in covariates. We analyzed 11 contexts, including binary contexts (sex)

and continuous contexts (top two PCs, age, BMI, smoking, alcohol, employment, education, income, number of years living in current address).

Population descriptor usage. We explain our usage choices of population descriptor, including the use of top two PCs to capture genetic ancestry/similarity and the use of "white British" in analyses of UK Biobank and "white SIRE" in analyses of All of Us. We use the top two PCs computed across all individuals in UK Biobank or in All of Us, respectively, to capture the continuous genetic ancestry variation in each dataset. While these two PCs provide major axes of genetic variation (Supplementary Fig. 3), we acknowledge that top two PCs alone are not sufficient to fully capture all variation in the entire population. We used discretized PC1 and PC2 subgroups to calculate population-level statistics such as R^2 while we acknowledge that the underlying genetic variation is continuous. In UK Biobank, we intended to analyze a set of individuals with relatively similar genetic ancestry to perform GWAS and derive PGS. We used a set of individuals previously annotated with "white British" that were identified using a combination of self-reported ethnic background and genetic information having very similar ancestral backgrounds based on PCA results[28]. In All of Us, we selected a set of individuals with self-reported race/ethnicity (SIRE) being "white", to study how PGS have different accuracy across environmental contexts in such a sample defined by SIRE. Noting that SIRE is not equivalent to genetic ancestry, the contrast of results from UK Biobank and All of Us helps understand how genetic and non-genetic factors impact PGS accuracy in a group of individuals defined by SIRE or genetic ancestry.

Evaluating context-specific prediction intervals. For quantitative traits, noting that prediction mean and standard deviation are $\hat{\mu}(\mathbf{c}), \hat{\sigma}(\mathbf{c})$ for a target individual with contexts $\mathbf{c}$, we evaluate prediction intervals with regard to phenotypes y using metrics of (1) prediction accuracy:

$R^2(\hat{\mu}(\mathbf{c}), y)$; (2) coverage of prediction intervals: evaluating $\Pr\left\{y \in \left[\hat{\mu}(\boldsymbol{c}_i) - \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\hat{\sigma}(\boldsymbol{c}_i), \hat{\mu}(\boldsymbol{c}_i) + \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\hat{\sigma}(\boldsymbol{c}_i)\right]\right\} \approx \alpha$, i.e., whether prediction intervals cover true phenotypes with pre-specified probability of α. Both metrics are evaluated both across all individuals, and within each context subgroup. We generated and evaluated context-specific intervals in both UK Biobank and All of Us. Prediction mean includes predictors of PGS, age, age*sex, age^2, top 10 PCs, and contexts in Fig. 2 and 3. Prediction variance includes predictors of age, sex, PC1, PC2, and contexts in Fig. 2 and 3. For each trait, we performed evaluation by repeatedly randomly sampling 5,000 individuals as calibration data, and 5,000 individuals as target data (as described in "Constructing calibrated and context-specific intervals").

Evaluating disease probability predictions. For disease traits, denoting binary disease status as y and predicted probability as $\hat{p}(\mathbf{c})$, we evaluate calibration of disease probability. For each predicted probability bin $[p_{\text{low}}, p_{\text{high}}]$, we examine whether the observed disease prevalence $P\left[y = 1 | \hat{p}(\mathbf{c}) \in [p_{\text{low}}, p_{\text{high}}]\right]$ is approximately equal to $\frac{p_{\text{low}} + p_{\text{high}}}{2}$. Calibration is evaluated for all individuals, and for each context subgroup.

We analyzed four well-powered disease trait GWAS in All of Us: type 2 diabetes (T2D)[50,61], coronary artery disease (CAD)[51], prostate cancer (PrCA)[52], breast cancer (BrCA)[53]. We predict disease probability using four models by incrementally adding complexity. (1) "Baseline": logistic regression using PGS, age, age^2, sex, age*sex, top 10 PCs, BMI, BMI^2 as predictors; (2) "Baseline+C": with additional context predictors of smoking, alcohol, employment, education, income, number of years living in current address; (3) "Baseline+C+PGSxC": with additional PGSxContext interaction terms; (4) "Baseline+C+PGSxC (VbyC)": with additional context-specific variance as a function of contexts.

Simulation studies of context-specific calibration

We performed simulations for both quantitative and disease traits with gene-context interactions.

Simulations of quantitative traits with gene-context interactions. For quantitative traits, we evaluated CalPred under three common scenarios of gene-context interactions in two contexts. Denoting genetic and environmental components in two contexts as G_1, G_2, E_1, E_2, these three scenarios include: (1) imperfect genetic correlation: $\mathrm{Cor}[G_1, G_2] < 1$, $\mathrm{Var}[G_1] = \mathrm{Var}[G_2]$ and $\mathrm{Var}[E_1] = \mathrm{Var}[E_2]$; (2) varying genetic variance: $\mathrm{Cor}[G_1, G_2] = 1$, $\mathrm{Var}[G_1] \neq \mathrm{Var}[G_2]$ and $\mathrm{Var}[E_1] = \mathrm{Var}[E_2]$; (3) proportional amplification of genetic and environmental components: $\mathrm{Cor}[G_1, G_2] = 1$, $\mathrm{Var}[G_1] \neq \mathrm{Var}[G_2]$, $\mathrm{Var}[E_1] \neq \mathrm{Var}[E_2]$, while ratios between G and E are the same across contexts: $\frac{\mathrm{Var}[G_1]}{\mathrm{Var}[E_1]} = \frac{\mathrm{Var}[G_2]}{\mathrm{Var}[E_2]}$. Across three scenarios, PGS weights derived in the first context were applied in both contexts. We evaluated the bias in prediction mean and coverage of prediction intervals.

Simulations of disease traits with gene-context interactions. For disease traits, we performed simulations with gene-context interactions in two contexts under a liability threshold model. These three scenarios include: (1) imperfect genetic correlation; (2) varying genetic variance; (3) varying disease prevalence where G_1, E_1 and G_2, E_2 are simulated using the same model but the disease prevalence is different across contexts. PGS weights derived in first context were applied to individuals in both contexts. We fit four regression models using different sets of predictors across all individuals: (1) y ~ PGS; (2) y ~ PGS + Context; (3) y ~ PGS + PGSxContext; (4) y ~ PGS + PGSxContext + Context. We note that logistic and probit regression models produced similar results.

Simulations of quantitative traits with multiple contexts. We simulated PGS point predictions $\hat{y}$ and phenotype values y to simulate traits with variable prediction accuracy across

genetic ancestry, age, and sex. We started with real contexts from UK Biobank individuals not used for PGS training (see section "Real data analyses"). We quantile normalized each context so they had mean 0 and variance 1. Such simulations preserved the correlation between contexts. Given these processed contexts, we simulated point predictions $\hat{y}$ using a normal distribution $\hat{y} \sim \mathcal{N}(0,1)$, and we simulated phenotypes y with:

$$y \sim \mathcal{N}(\hat{y}, \exp\left(\beta_{\sigma,0} + \sum_{c} \beta_{\sigma,c} \times c\right),$$

where $\beta_{\sigma,0}$ denoted the baseline variance of y, and $\beta_{\sigma,c}$ was the effect of context c to the variance of y. "Σ_c" enumerated over PC1, age, sex. This procedure simulated different variance of y around $\hat{y}$ for individuals with different contexts, as observed in real data. We first selected $\beta_{\sigma,0}$ such that $R^2(y, \hat{y}) = 30\%$ for individuals with average contexts ($\sum_c \beta_{\sigma,c} \times c = 0$). We simulated data with variable variances: we set $\beta_{\sigma,\text{age}} = 0.25, \beta_{\sigma,\text{sex}} = 0.2, \beta_{\sigma,\text{PC1}} = 0.15$. These parameters were manually chosen to match observed variable R^2 in real data. In each simulation, we randomly sampled N_{cal}=100, 500, 2500, 5000 individuals used for estimating the calibration model and N_{test} = 5000 individuals for evaluation. New point predictions and phenotypes $\hat{y}, y$ were simulated in each simulation. And we quantified prediction accuracy and coverage of prediction intervals in each simulation replicate.

Statistics and reproducibility

We analyzed two publicly available datasets of UK Biobank and All of Us, and sample sizes were determined in these studies. We did not use randomization or blinding. No data was excluded from the analyses. We replicated our findings across these two independent datasets.

4.5 Figures

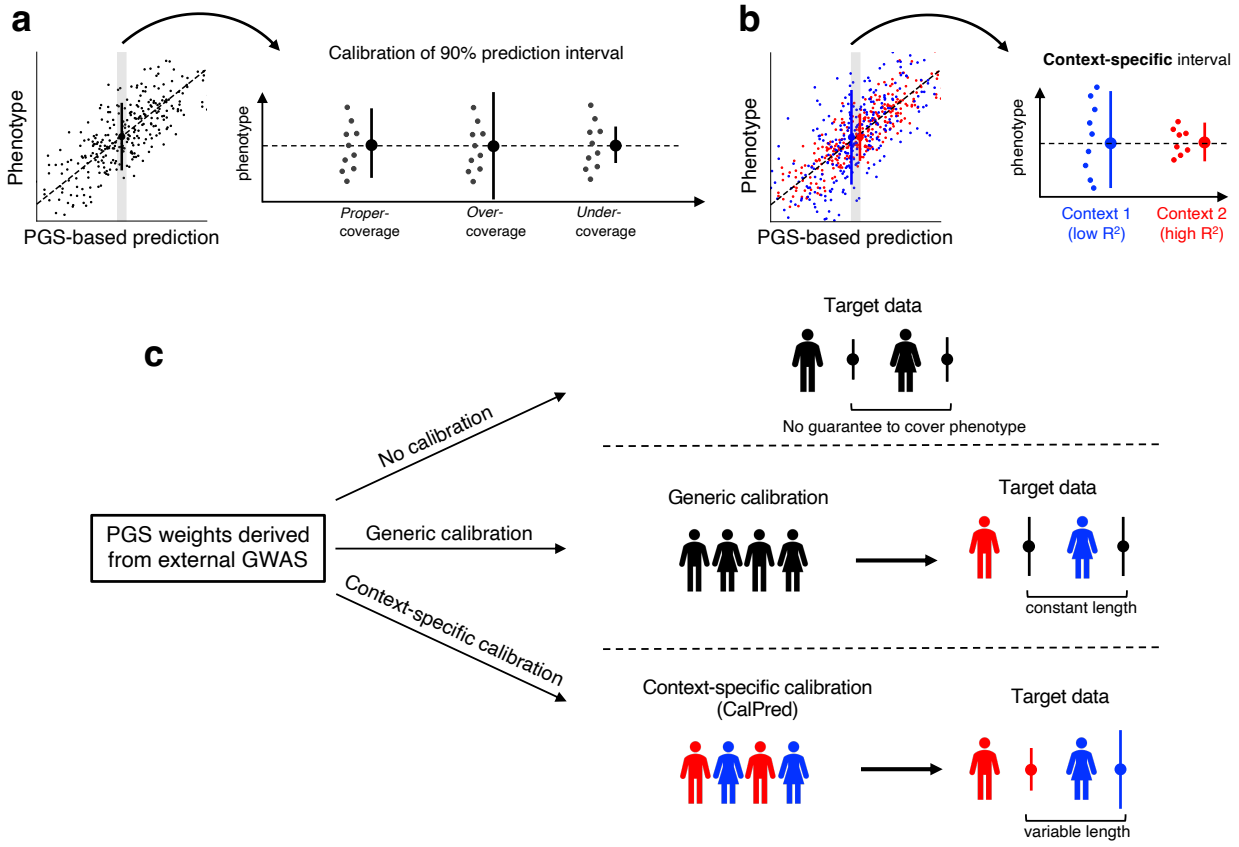

Figure 4.1 Calibrated and context-specific prediction intervals via CalPred.

(a) Calibration of prediction intervals. We consider a set of individuals with the same point prediction (shaded area in the left panel, dashed horizontal line in the right panel). Each dot denotes an individual's phenotype value. Intervals with *proper-coverage* cover the true phenotype at pre-specified probability of 90%; intervals with *over-coverage* are incorrectly wide; intervals with *under-coverage* are incorrectly narrow. **(b)** Context-specific calibration of prediction intervals. We consider two subpopulations in different contexts (e.g., female and male). Context 1 has lower prediction accuracy

and therefore wider variation around the mean, while context 2 has higher prediction accuracy and therefore narrower variation around the mean. Context-specific intervals vary by context, providing intervals with proper coverage in each context. **(c)** Different approaches for prediction intervals of PGS-based models. All approaches start with a set of predefined PGS weights derived from existing GWAS. “No calibration”: prediction intervals can be calculated using analytical formula without calibration data. However, these intervals are not guaranteed to be well-calibrated. “Generic calibration”: these methods do not consider context information; they produce generic prediction intervals that are constant across individuals. “Context-specific calibration”: these methods leverage a set of calibration data to estimate the impact of each context to trait prediction accuracy; the estimated impact can then be used to generate prediction intervals for any target individuals matching in distribution with calibration data.

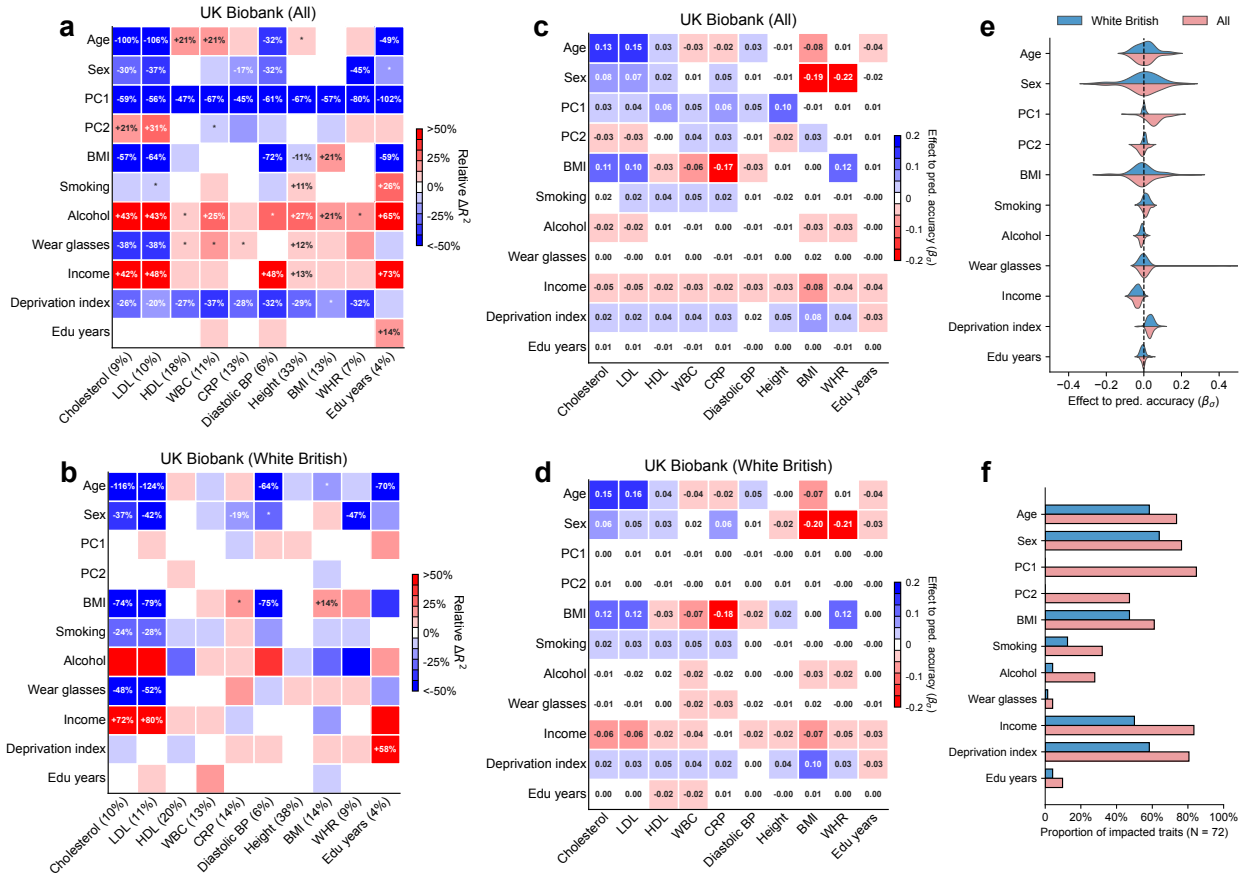

Figure 4.2 Widespread context-specific PGS prediction accuracy in UK Biobank.

(a-b) Heatmaps for context-specific PGS accuracy for all and WB individuals. Each row denotes a context and each column denotes a trait; the squared correlation between PGS and residual phenotype (R^2) is shown in parentheses. Heatmap color denotes the PGS-phenotype relative ΔR^2 (defined as $\frac{R^2_{\text{group1}} - R^2_{\text{group2}}}{R^2_{\text{all}}}$), where $R^2_{[\text{subset}]}$ represents R^2 computed in a given range of the context variable. For continuous contexts, relative ΔR^2 denote differences of top quintile minus bottom quintile; for binary contexts (including sex, smoking, wear glasses, alcohol), relative ΔR^2 denote differences of male minus female, smoking minus not smoking, wearing glasses minus not wearing glasses, drinking alcohol minus not drinking alcohol (these orders were arbitrarily chosen). Numerical values of relative

R^2 differences are displayed for PGS-context pairs with statistically significant differences (multiple testing correction for all 10×11 PGS-context pairs in this figure; two-sided $p < 0.05 / (10\times11)$). '*' are displayed for PGS-context pairs with nominally significant differences (multiple testing correction for 11 contexts; two-sided $p < 0.05 / 11$). **(c-d)** Heatmaps of effects to prediction accuracy in CalPred model (estimated β_σ). Colormaps were inversed to those of **(a-b)** to reflect that positive β_σ corresponds to lower prediction accuracy and vice versa. **(e)** Distribution of estimated β_σ in the CalPred model for each context across traits. **(f)** Number of significantly impacted traits by each context (two-sided $p < 0.05 / (72\times11)$).

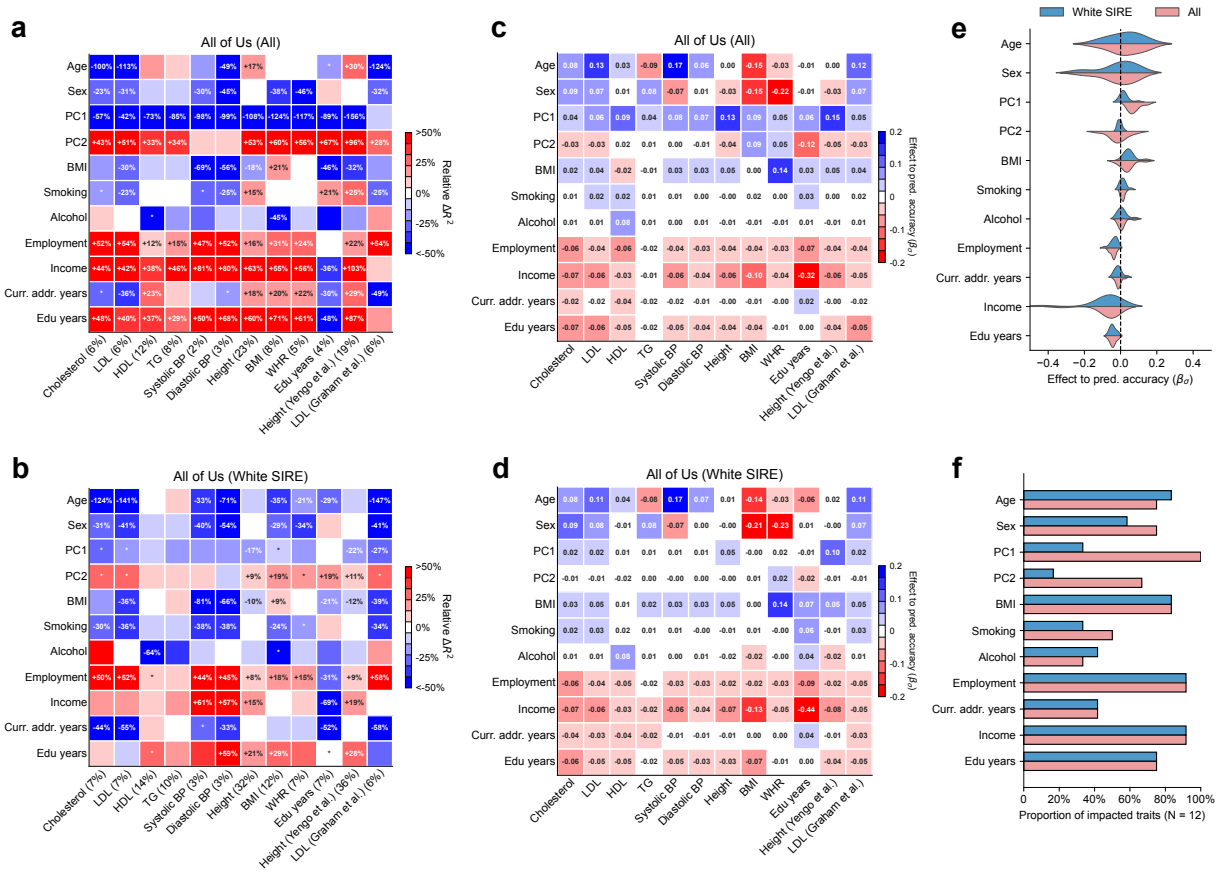

Figure 4.3 Widespread context-specific PGS prediction accuracy in All of Us.

(a-b) Heatmaps for context-specific PGS accuracy for all and white SIRE individuals. Each row denotes a context and each column denotes a trait; overall R^2 is shown in parentheses. Heatmap color denotes relative ΔR^2: differences of top quintile minus bottom quintile for continuous contexts and difference of male minus female for binary context of sex. Numerical values of relative R^2 differences are displayed for trait-context pairs with statistically significant differences (multiple testing correction for all 12×11 PGS-context pairs in this figure; two-sided $p < 0.05 / (12\times11)$). '*' are displayed for PGS-context pairs with nominally significant differences (multiple testing correction for 11 contexts; two-sided $p < 0.05 / 11$). **(c-d)** Heatmaps of estimated β_σ in CalPred model. **(e)** Distribution of estimated β_σ in CalPred model for each context across traits. **(f)** Number of significantly impacted traits by each context (with two-sided $p < 0.05 / (12\times11)$).

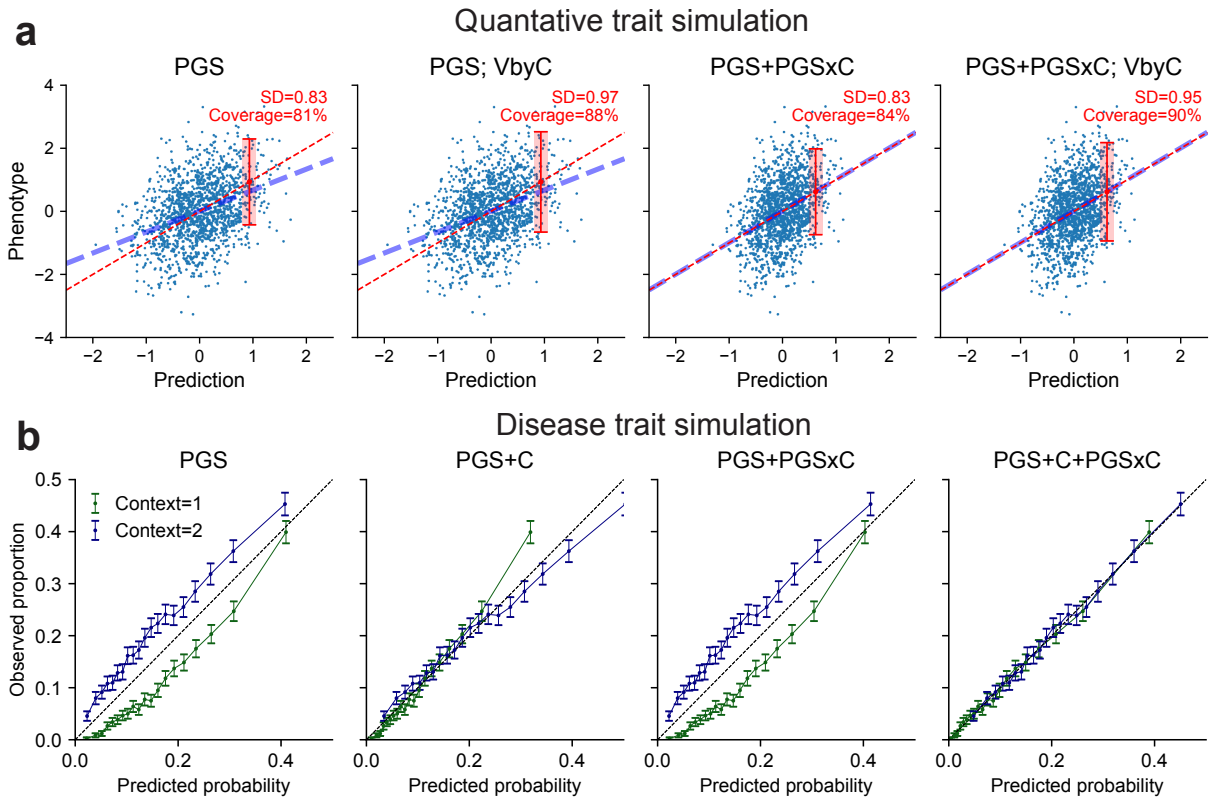

Figure 4.4 Simulation studies with gene-context interactions.

(a) quantitative traits. We simulated traits for individuals in two contexts with 0.7 cross-context genetic correlation and heritability=0.5/0.4 respectively in two contexts. PGS weights were trained in the first context and applied in the second context. We showed results for predictions in the second context using four combinations of approaches to model prediction mean (using PGS or PGS+PGSxC) and prediction variance (with or without modeling variance by contexts; VbyC). We did not simulate effects of context variables to phenotype and therefore results using "PGS+C" and "PGS+C+PGSxC" would yield same results as "PGS" and therefore not included. Blue dashed line denotes the best fit to data; red dashed line denotes model predictions; red error bar denotes the CalPred 90% prediction interval for individual at top 5% quantile of PGS. Prediction interval coverage was evaluated within data in top PGS decile. Additional details can be found in Extended Data Fig. 5 and Methods. **(b) disease traits.** We simulated diseases for individuals in two contexts under a liability threshold model with 0.5 heritability, 0.7 cross-context genetic correlation and disease prevalence=10%/20% respectively in two contexts. Disease probability was predicted using four sets of predictors: (1) PGS; (2) PGS and context variables (PGS+C); (3) PGS and PGSxContext interactions (PGS+PGSxC); (4) PGS, context variables and PGSxContext interactions (PGS+C+PGSxC). Modeling variance as function of contexts (VbyC) led to similar results. Error bars denote observed disease proportions and their 95% confidence intervals for each predicted probability bin (n=2000 individuals for each error bar). Additional details can be found in Extended Data Fig. 6 and Methods.

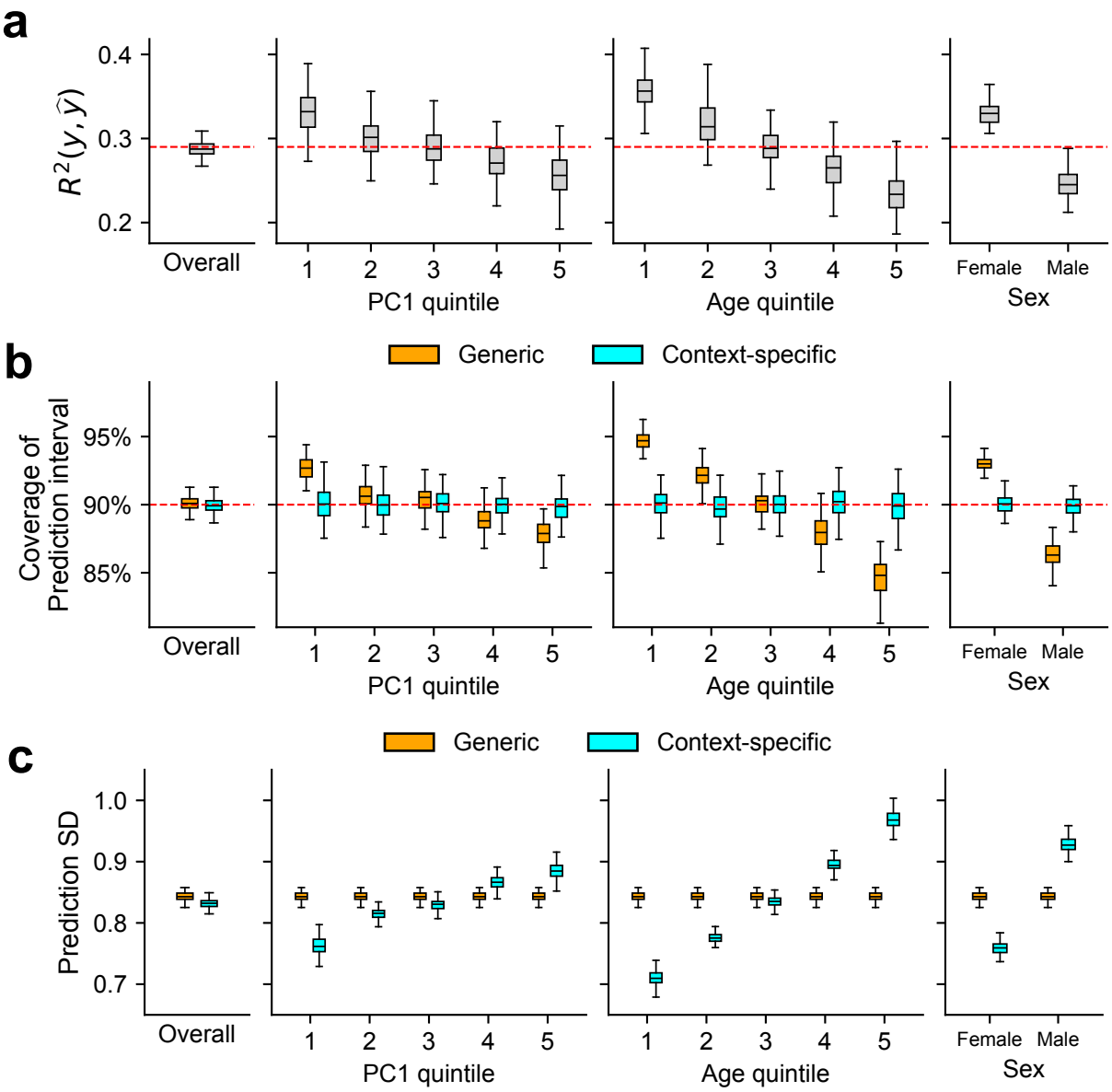

Figure 4.5 Simulation studies with multiple contexts.

Simulations were performed to reflect scenarios where individuals have variable prediction accuracy by genetic PC1, age, and sex. For each simulation, we first trained a calibration model using a random set of 5,000 training individuals and then evaluated resulting prediction intervals on 5,000 target individuals (Methods). **(a)** Prediction R^2 between y and $\hat{y}$ in simulated data both at the overall level,

and in each context subgroup. **(b)** Coverage of generic vs. context-specific 90% prediction intervals evaluated in each context subgroup. Generic intervals were obtained by applying CalPred without context information; context-specific intervals were obtained by applying CalPred together with context information. **(c)** Average length of generic vs. context-specific prediction standard deviation (SD) in each context. Each box plot contains R^2/coverage/average length evaluated across 100 simulations (n=100 points for each box), the center corresponds to the median; the box represents the first and third quartiles of the points; the whiskers represent the minimum and maximum points located within 1.5× interquartile range from the first and third quartiles, respectively.

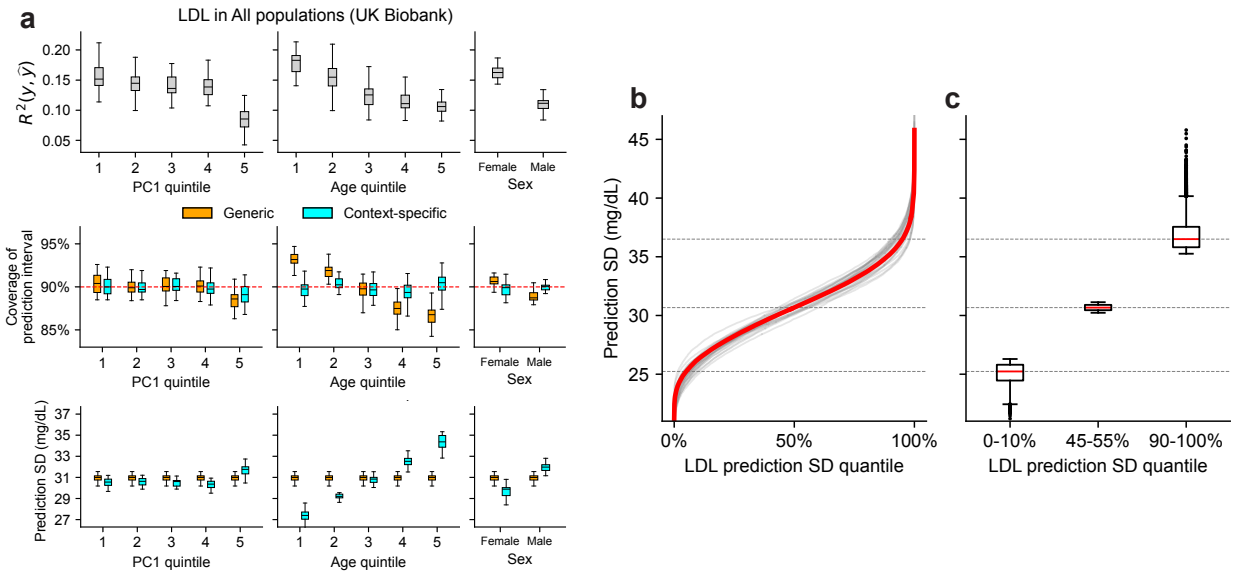

Figure 4.6 CalPred PGS calibration of LDL in UK Biobank.

(a) Top panel: Prediction R^2 between phenotype and point predictions (incorporating PGS and other covariates) in each subgroup of individuals stratified by context (R^2 evaluated across all individuals is 0.147). **Middle panel:** Coverage of generic vs. context-specific 90% prediction intervals evaluated in

each context subgroup. Generic intervals were obtained by applying CalPred without context information; context-specific intervals were obtained by applying CalPred together with context information. **Bottom panel:** Average length of generic vs. context-specific 90% prediction intervals in each context. Each box plot contains R^2/coverage/average length across 30 random samples with each sample of 5,000 training and 5,000 target individuals (n=30 points for each box). **(b)** Ordered LDL prediction SD in unit of mg/dL. Gray lines denote prediction SD obtained with random sample of 5,000 training and applied to 5,000 target individuals. Red line denote prediction SD obtained from all individuals. **(c)** Box plots of results in (b) from individuals of LDL prediction SD quantile of 0-10%, 45-55%, 90-100% (n=110K individuals in total). For box plots in both (a) and (c), the center corresponds to the median; the box represents the first and third quartiles of the points; the whiskers represent the minimum and maximum points located within 1.5× interquartile ranges from the first and third quartiles, respectively.

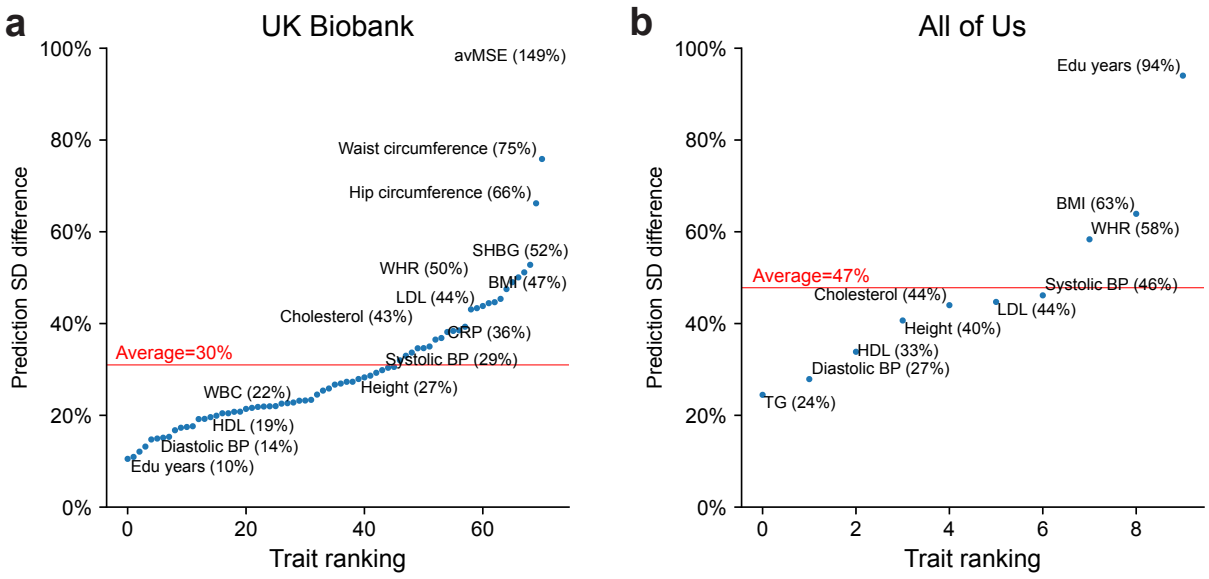

Figure 4.7 Variation of prediction standard deviation (SD) accounting for all contexts.

Relative difference of prediction SD between top and bottom prediction SD deciles (90-100% vs. 0-10%) for all traits in UK Biobank **(a)** and All of Us **(b)**. Traits are ranked by prediction SD. The difference is calculated with the median prediction SD within decile of individuals with highest prediction SD s_{d1} and decile of individuals with lowest prediction SD s_{d10} using $\left(\frac{s_{d1}-s_{d10}}{s_{d10}}-1\right)\times 100\%$. avMSE, average mean spherical equivalent; SHBG, sex hormone binding globulin; LDL, low-density lipoprotein cholesterol; WHR, waist-hip ratio; BMI, body mass index; CRP, C-reactive protein; BP, blood pressure; WBC, white blood cell count; HDL, high-density lipoprotein cholesterol; TG, triglycerides.

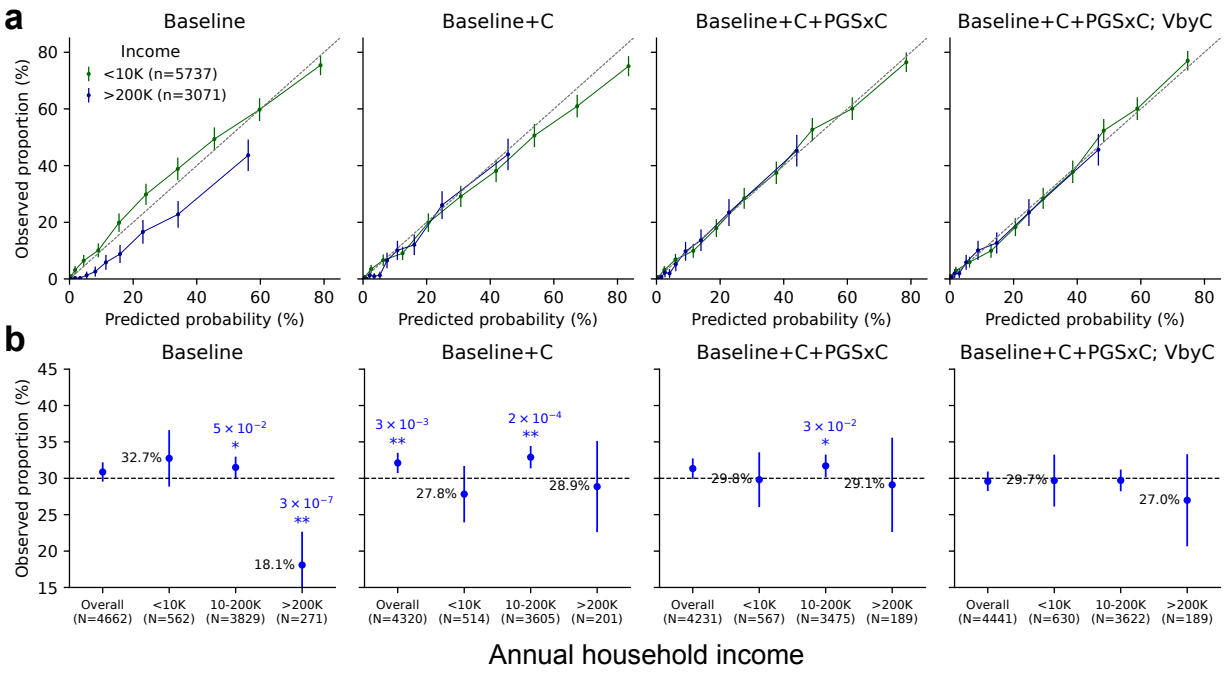

Figure 4.8 Calibration of T2D risk prediction across income groups.

We compared four models for predicting T2D across all individuals in All of Us. “Baseline”: logistic regression model with PGS, age, sex, BMI, top 10 PCs as predictors; “Baseline+C”: logistic regression model additionally including smoking status, drinking, employment, income, current address years and education; “Baseline+C+PGSxC”: additionally include PGSxC interactions; “Baseline+C+PGSxC;

VbyC": additionally modeling variance by contexts within a liability threshold model. The dataset was evenly split into training and testing datasets. **(a)** Observed proportion versus predicted probability of T2D for lowest and highest income groups. Error bars denote the observed proportions and their 95% confidence intervals (number of total individuals shown in legend). **(b)** Observed proportion of individuals with T2D among individuals predicted with a predicted T2D risk of approximately 30% (25%-35%) for baseline and calibrated models stratified by annual household income. Error bars denote the observed proportions and their 95% confidence intervals (number of individuals for each error bar is shown in parenthesis). '*' and '**' denote statistical significance levels for deviations from the 30% predicted risk, with '*' indicating $p<0.05$ and '**' denoting $p<0.01$, respectively (two-sided tests).

5 Conclusion

This dissertation has introduced new methodologies for genetic mapping, inference and prediction across diverse human populations. Chapter 2 highlights the importance of selecting appropriate statistical procedures for genetic mapping in individuals of diverse genetic ancestry to maximize discovery power; Chapter 3 demonstrates that genetic effects are shared across genetic ancestry backgrounds when environmental confounding is properly accounted for; Chapter 4 characterizes the pervasive context-specific accuracy of polygenic scoring methods and shows that accounting for such will improve portability across contexts. Together, these findings form a cohesive foundation for the field to develop genetically-informed precision medicine that benefits everyone.

Chapter 2 and 3 collectively suggest that the biological effects of genetic variation are largely consistent across ancestry backgrounds. This finding has significant implications for future sample collection strategies in genetic studies. Including genetic samples from individuals of diverse genetic ancestry can more efficiently capture additional genetic variants that are rare in European but common in other genetic ancestry groups, leading to a higher yield of new discoveries within the same research budget[141]. Of particular interest are individuals of admixed ancestry, who inherit genetic variants from multiple ancestral populations. As the biological effects are shared across populations, drugs and treatments developed based on data from one ancestry group are likely to portable to other populations, even though the corresponding genetic variants have different frequencies across populations.

In Chapter 4, I have highlighted the widespread differential performance of polygenic scores across context groups. Such context-specific accuracy of PGS is highly pervasive across traits and biobanks; socioeconomic contexts often having larger impact than genetic ancestry. I

introduce a new approach, CalPred, to estimate context-specific prediction intervals, providing a framework to quantify individualized context-specific generalizability/portability of a given polygenic score. CalPred equips practitioners with a personalized metric of reliability and a clear understanding of associated confidence levels for each prediction, aiding the application and responsible deployment of polygenic scores across diverse cohorts. Furthermore, this work highlights the importance of comprehensive profiling of context information in future studies to account for and mitigate the portability issues associated with polygenic scores.

Genetics and genomics research has made rapid progress over the past decades and has great potential to be implemented our medical and clinical practices. In realizing such transition, two major challenges have emerged that need to be addressed. First, how to account for the diverse backgrounds in both genetic ancestry and context distribution across human populations? Second, how to integrate the vast amount of information available for each individual, together with their genetic and genomics data? The approaches introduced here can be applied to the growing amount of genetic data from diverse populations to tackle these challenges. By leveraging diversity in genetic and context backgrounds, these methods improve discovery power (Chapter 2), dissect the level of genetic effects sharing across ancestry backgrounds while accounting for environmental confounding (Chapter 3), and can be applied across all individuals incorporating the differential performance of polygenic scores across contexts for predictive analyses (Chapter 4). These methodologies serve as a foundation for building a more inclusive and effective framework for precision medicine that can improve health outcomes for all individuals, regardless of their genetic ancestry or socioeconomic background.

Moving forward, I anticipate significant advancements in incorporating additional information sources and developing computational models that directly interface with clinical applications.

This will involve fully exploiting personal health status measurements, including rare genetic variants, multi-omics data (e.g., transcriptomics, proteomics, and metabolomics) and comprehensive environmental factor profiling (e.g., lifestyle, diet, and exposure to pollutants). Integration of these diverse data modalities, together with new analytical tools that accurately model their additive and interaction effects, will provide a holistic view of an individual's health, enabling accurate risk assessments, personalized treatment strategies, and targeted preventive measures. For the successful transition of genetic research into clinical application, it is important to identify clinically relevant evaluation metrics and designing simple, implementable measures for informed decision-making based on a patient's genetic profile and other relevant health information. To conclude, I am hopeful that the increasing data availability from diverse populations and modalities, coupled with the development of novel analytical approaches, such as those presented in this thesis, will pave the way for precision medicine with more personalized and effective healthcare solutions.

6 References

1. Polderman, T. J. C. *et al.* Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nat. Genet.* **47**, 702–709 (2015).

2. Hunter, D. J. Gene–environment interactions in human diseases. *Nat. Rev. Genet.* **6**, 287–298 (2005).

3. Kris A. Wetterstrand, M. S. DNA sequencing costs: Data. *Genome.gov* https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data (2019).

4. Ashley, E. A. Towards precision medicine. *Nat. Rev. Genet.* **17**, 507–522 (2016).

5. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five years of GWAS discovery. *Am. J. Hum. Genet.* **90**, 7–24 (2012).

6. Visscher, P. M. *et al.* 10 years of GWAS discovery: Biology, function, and translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).

7. Abdellaoui, A., Yengo, L., Verweij, K. J. H. & Visscher, P. M. 15 years of GWAS discovery: Realizing the promise. *Am. J. Hum. Genet.* (2023) doi:10.1016/j.ajhg.2022.12.011.

8. Menzel, S. *et al.* A QTL influencing F cell production maps to a gene encoding a zinc-finger protein on chromosome 2p15. *Nat. Genet.* **39**, 1197–1199 (2007).

9. Uda, M. *et al.* Genome-wide association study shows *BCL11A* associated with persistent fetal hemoglobin and amelioration of the phenotype of β-thalassemia. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 1620–1625 (2008).

10. Nelson, M. R. *et al.* The support of human genetic evidence for approved drug indications. *Nat. Genet.* **47**, 856–860 (2015).

11. Minikel, E. V., Painter, J. L., Dong, C. C. & Nelson, M. R. Refining the impact of genetic evidence on clinical success. *Nature* 1–6 (2024).

12. Khera, A. V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* **50**, 1219–1224 (2018).

13. Natarajan, P. *et al.* Polygenic risk score identifies subgroup with higher burden of atherosclerosis and greater relative benefit from statin therapy in the primary prevention setting. *Circulation* **135**, 2091–2101 (2017).

14. Kachuri, L. *et al.* Genetically adjusted PSA levels for prostate cancer screening. *Nat. Med.* **29**, 1412–1423 (2023).

15. Mills, M. C. & Rahal, C. The GWAS Diversity Monitor tracks diversity by disease in real time. *Nat. Genet.* **52**, 242–243 (2020).

16. Wojcik, G. L. *et al.* Genetic analyses of diverse populations improves discovery for complex traits. *Nature* **570**, 514–518 (2019).

17. The SIGMA Type 2 Diabetes Consortium. Sequence variants in SLC16A11 are a common risk factor for type 2 diabetes in Mexico. *Nature* **506**, 97–101 (2014).

18. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).

19. Martin, A. R. *et al.* Human demographic history impacts genetic risk prediction across diverse populations. *Am. J. Hum. Genet.* **107**, 788–789 (2020).

20. Ding, Y. *et al.* Polygenic scoring accuracy varies across the genetic ancestry continuum. *Nature* **618**, 774–781 (2023).

21. Mostafavi, H. *et al.* Variable prediction accuracy of polygenic scores within an ancestry group. *Elife* **9**, e48376 (2020).

22. Zhou, W. *et al.* Global Biobank Meta-analysis Initiative: Powering genetic discovery across human disease. *Cell Genomics* **2**, 100192 (2022).

23. Johnson, R. *et al.* The UCLA ATLAS Community Health Initiative: Promoting precision health research in a diverse biobank. *Cell Genom.* **3**, 100243 (2023).

24. The All of Us Research Program Genomics Investigators *et al.* Genomic data in the All of Us Research Program. *Nature* **627**, 340–346 (2024).

25. Hou, K., Bhattacharya, A., Mester, R., Burch, K. S. & Pasaniuc, B. On powerful GWAS in admixed populations. *Nature genetics* vol. 53 1631–1633 (2021).

26. Hou, K. *et al.* Causal effects on complex traits are similar for common variants across segments of different continental ancestries within admixed individuals. *Nat. Genet.* **55**, 549–558 (2023).

27. Hou, K., Xu, Z., Ding, Y., Harpak, A. & Pasaniuc, B. Calibrated prediction intervals for polygenic scores across diverse contexts. *medRxiv* (2023) doi:10.1101/2023.07.24.23293056.

28. Martin, A. R. *et al.* Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *Am. J. Hum. Genet.* **100**, 635–649 (2017).

29. Atkinson, E. G. *et al.* Tractor uses local ancestry to enable the inclusion of admixed individuals in GWAS and to boost power. *Nat. Genet.* (2021) doi:10.1038/s41588-020-00766-y.

30. Zhang, J. & Stram, D. O. The role of local ancestry adjustment in association studies using admixed populations. *Genet. Epidemiol.* **38**, 502–515 (2014).

31. Seldin, M. F., Pasaniuc, B. & Price, A. L. New approaches to disease mapping in admixed populations. *Nat. Rev. Genet.* **12**, 523–528 (2011).

32. Pasaniuc, B. *et al.* Enhanced statistical tests for GWAS in admixed populations: assessment using African Americans from CARe and a Breast Cancer Consortium. *PLoS Genet.* **7**, e1001371 (2011).

33. Tang, H., Siegmund, D. O., Johnson, N. A., Romieu, I. & London, S. J. Joint testing of genotype and ancestry association in admixed families. *Genetic Epidemiology* vol. 34 783–791 Preprint at https://doi.org/10.1002/gepi.20520 (2010).

34. Shriner, D., Adeyemo, A. & Rotimi, C. N. Joint ancestry and association testing in admixed individuals. *PLoS Comput. Biol.* **7**, e1002325 (2011).

35. Yorgov, D., Edwards, K. L. & Santorico, S. A. Use of admixture and association for detection of quantitative trait loci in the Type 2 Diabetes Genetic Exploration by Next-Generation Sequencing in Ethnic Samples (T2D-GENES) study. *BMC Proceedings* vol. 8 Preprint at https://doi.org/10.1186/1753-6561-8-s1-s6 (2014).

36. Wojcik, G. L. *et al.* Genetic analyses of diverse populations improves discovery for complex traits. *Nature* **570**, 514–518 (2019).

37. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

38. de Candia, T. R. *et al.* Additive Genetic Variation in Schizophrenia Risk Is Shared by Populations of African and European Descent. *Am. J. Hum. Genet.* **93**, 463–470 (2013).

39. Lam, M. *et al.* Comparative genetic architectures of schizophrenia in East Asian and European populations. *Nat. Genet.* **51**, 1670–1678 (2019).

40. Liu, J. Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* **47**, 979–986 (2015).

41. Shi, H. *et al.* Population-specific causal disease effect sizes in functionally important regions impacted by selection. *Cold Spring Harbor Laboratory* 803452 (2019) doi:10.1101/803452.

42. Van Rheenen, W., Peyrot, W. J., Schork, A. J., Lee, S. H. & Wray, N. R. Genetic correlations of polygenic disease traits: from theory to practice. *Nat. Rev. Genet.* **20**, 567–581 (2019).

43. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

44. Ramirez, A. H. *et al.* The All of Us Research Program: Data quality, utility, and diversity. *Patterns (N Y)* **3**, 100570 (2022).

45. Brown, B. C., Ye, C. J., Price, A. L. & Zaitlen, N. Transethnic genetic-correlation estimates from summary statistics. *Am. J. Hum. Genet.* **99**, 76–88 (2016).

46. Galinsky, K. J. *et al.* Estimating cross-population genetic correlations of causal effect sizes. *Genet. Epidemiol.* **43**, 180–188 (2019).

47. Shi, H. *et al.* Localizing components of shared transethnic genetic architecture of complex traits from GWAS summary data. *Am. J. Hum. Genet.* **106**, 805–817 (2020).

48. Shi, H. *et al.* Population-specific causal disease effect sizes in functionally important regions impacted by selection. *Nat. Commun.* **12**, 1098 (2021).

49. Kanai, M. *et al.* Insights from complex trait fine-mapping across diverse populations. *bioRxiv* (2021) doi:10.1101/2021.09.03.21262975.

50. Wang, Y. *et al.* Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations. *Nat. Commun.* **11**, 3865 (2020).

51. Gurdasani, D., Barroso, I., Zeggini, E. & Sandhu, M. S. Genomics of disease risk in globally diverse populations. *Nat. Rev. Genet.* **20**, 520–535 (2019).

52. Sirugo, G., Williams, S. M. & Tishkoff, S. A. The missing diversity in human genetic studies. *Cell* **177**, 1080 (2019).

53. Marigorta, U. M. & Navarro, A. High trans-ethnic replicability of GWAS results implies common causal variants. *PLoS Genet.* **9**, e1003566 (2013).

54. Patel, R. A. *et al.* Genetic interactions drive heterogeneity in causal variant effect sizes for gene expression and complex traits. *Am. J. Hum. Genet.* **109**, 1286–1297 (2022).

55. Cai, N. *et al.* Minimal phenotyping yields genome-wide association signals of low specificity for major depression. *Nat. Genet.* **52**, 437–447 (2020).

56. Bitarello, B. D. & Mathieson, I. Polygenic Scores for Height in Admixed Populations. *G3* **10**, 4027–4036 (2020).

57. Marnetto, D. *et al.* Ancestry deconvolution and partial polygenic score can improve susceptibility predictions in recently admixed individuals. *Nat. Commun.* **11**, 1628 (2020).

58. Bentley, A. R. *et al.* Gene-based sequencing identifies lipid-influencing variants with ethnicity-specific effects in African Americans. *PLoS Genet.* **10**, e1004190 (2014).

59. Rajabli, F. *et al.* Ancestral origin of ApoE ε4 Alzheimer disease risk in Puerto Rican and African American populations. *PLoS Genet.* **14**, e1007791 (2018).

60. Blue, E. E., Horimoto, A. R. V. R., Mukherjee, S., Wijsman, E. M. & Thornton, T. A. Local ancestry at APOE modifies Alzheimer's disease risk in Caribbean Hispanics. *Alzheimers. Dement.* **15**, 1524–1532 (2019).

61. Naslavsky, M. S. *et al.* Global and local ancestry modulate APOE association with Alzheimer's neuropathology and cognitive outcomes in an admixed sample. *Mol. Psychiatry* **27**, 4800–4808 (2022).

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

63. Sakaue, S. *et al.* A cross-population atlas of genetic associations for 220 human phenotypes. *Nat. Genet.* **53**, 1415–1424 (2021).

64. Zeng, J. *et al.* Signatures of negative selection in the genetic architecture of human complex traits. *Nat. Genet.* **50**, 746–753 (2018).

65. Schoech, A. P. *et al.* Quantification of frequency-dependent genetic architectures in 25 UK Biobank traits reveals action of negative selection. *Nat. Commun.* **10**, (2019).

66. Zhang, Y., Qi, G., Park, J.-H. & Chatterjee, N. Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits. *Nat. Genet.* **50**, 1318–1326 (2018).

67. The International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).

68. Speed, D. & Balding, D. J. SumHer better estimates the SNP heritability of complex traits from summary statistics. *Nat. Genet.* **51**, 277–284 (2019).

69. Deming, W. E. *Statistical Adjustment of Data*. (1964).

70. Hodonsky, C. J. *et al.* Ancestry-specific associations identified in genome-wide combined-phenotype study of red blood cell traits emphasize benefits of diversity in genomics. *BMC Genomics* **21**, 228 (2020).

71. Loh, P.-R. *et al.* Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nat. Genet.* **47**, 1385–1392 (2015).

72. Johnson, R. *et al.* Estimation of regional polygenicity from GWAS provides insights into the genetic architecture of complex traits. *PLoS Comput. Biol.* **17**, e1009483 (2021).

73. Liu, J., Lewinger, J. P., Gilliland, F. D., Gauderman, W. J. & Conti, D. V. Confounding and heterogeneity in genetic association studies with admixed populations. *Am. J. Epidemiol.* **177**, 351–360 (2013).

74. Saitou, M., Dahl, A., Wang, Q. & Liu, X. Allele frequency differences of causal variants have a major impact on low cross-ancestry portability of PRS. *bioRxiv* (2022) doi:10.1101/2022.10.21.22281371.

75. Pasaniuc, B. *et al.* Analysis of Latino populations from GALA and MEC studies reveals genomic loci with biased local ancestry estimation. *Bioinformatics* **29**, 1407–1415 (2013).

76. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).

77. Speed, D., Hemani, G., Johnson, M. R. & Balding, D. J. Improved heritability estimation from genome-wide SNPs. *Am. J. Hum. Genet.* **91**, 1011–1021 (2012).

78. Gazal, S. *et al.* Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. *Nat. Genet.* **49**, 1421–1427 (2017).

79. Gazal, S., Marquez-Luna, C., Finucane, H. K. & Price, A. L. Reconciling S-LDSC and LDAK functional enrichment estimates. *Nat. Genet.* **51**, 1202–1204 (2019).

80. Hou, K. *et al.* Accurate estimation of SNP-heritability from biobank-scale data irrespective of genetic architecture. *Nat. Genet.* **51**, 1244–1251 (2019).

81. Linnet, K. Performance of Deming regression analysis in case of misspecified analytical error ratio in method comparison studies. *Clin. Chem.* **44**, 1024–1031 (1998).

82. Chiu, A. M., Molloy, E. K., Tan, Z., Talwalkar, A. & Sankararaman, S. Inferring population structure in biobank-scale genomic data. *Am. J. Hum. Genet.* **109**, 727–737 (2022).

83. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).

84. Loh, P.-R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* **48**, 1443–1448 (2016).

85. Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.* **93**, 278–288 (2013).

86. The 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

87. Schubert, R., Andaleon, A. & Wheeler, H. E. Comparing local ancestry inference models in populations of two- and three-way admixture. *PeerJ* **8**, e10090 (2020).

88. Gay, N. R. *et al.* Impact of admixture and ancestry on eQTL analysis and GWAS colocalization in GTEx. *Genome Biol.* **21**, 233 (2020).

89. Pardiñas, A. F. *et al.* Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nat. Genet.* **50**, 381–389 (2018).

90. Schoech, A. P. *et al.* Negative short-range genomic autocorrelation of causal effects on human complex traits. *bioRxiv* 2020.09.23.310748 (2020) doi:10.1101/2020.09.23.310748.

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

92. Reich, D. *et al.* Reduced neutrophil count in people of African descent is due to a regulatory variant in the Duffy antigen receptor for chemokines gene. *PLoS Genet.* **5**, e1000360 (2009).

93. Reiner, A. P. *et al.* Genome-wide association and population genetic analysis of C-reactive protein in African American and Hispanic American women. *Am. J. Hum. Genet.* **91**, 502–512 (2012).

94. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).

95. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *J. R. Stat. Soc. Series B Stat. Methodol.* **82**, 1273–1300 (2020).

96. Chatterjee, N., Shi, J. & García-Closas, M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat. Rev. Genet.* **17**, 392–406 (2016).

97. Torkamani, A., Wineinger, N. E. & Topol, E. J. The personal and clinical utility of polygenic risk scores. *Nat. Rev. Genet.* **19**, 581–590 (2018).

98. Li, R., Chen, Y., Ritchie, M. D. & Moore, J. H. Electronic health records and polygenic risk scores for predicting disease risk. *Nat. Rev. Genet.* **21**, 493–502 (2020).

99. Kullo, I. J. *et al.* Polygenic scores in biomedical research. *Nat. Rev. Genet.* **23**, 524–532 (2022).

100. Privé, F. *et al.* Portability of 245 polygenic scores when derived from the UK Biobank and applied to 9 ancestry groups from the same cohort. *Am. J. Hum. Genet.* **109**, 373 (2022).

101. Weissbrod, O. *et al.* Leveraging fine-mapping and multipopulation training data to improve cross-population polygenic risk scores. *Nat. Genet.* **54**, 450–458 (2022).

102. Ruan, Y. *et al.* Improving polygenic prediction in ancestrally diverse populations. *Nat. Genet.* **54**, 573–580 (2022).

103. Mostafavi, H. *et al.* Variable prediction accuracy of polygenic scores within an ancestry group. *Elife* **9**, (2020).

104. Jiang, X., Holmes, C. & McVean, G. The impact of age on genetic risk for common diseases. *PLoS Genet.* **17**, e1009723 (2021).

105. Hui, D. *et al.* Quantifying factors that affect polygenic risk score performance across diverse ancestries and age groups for body mass index. *Pac. Symp. Biocomput.* **28**, 437–448 (2023).

106. Ding, Y. *et al.* Large uncertainty in individual polygenic risk score estimation impacts PRS-based risk stratification. *Nat. Genet.* **54**, 30–39 (2022).

107. Wray, N. R. *et al.* Pitfalls of predicting complex traits from SNPs. *Nat. Rev. Genet.* **14**, 507–515 (2013).

108. Ge, T., Chen, C.-Y., Neale, B. M., Sabuncu, M. R. & Smoller, J. W. Phenome-wide heritability analysis of the UK Biobank. *PLoS Genet.* **13**, e1006711 (2017).

109. Zhu, C. *et al.* Amplification is the primary mode of gene-by-sex interaction in complex human traits. *Cell Genom.* **3**, 100297 (2023).

110. Weine, E., Smith, S. P., Knowlton, R. K. & Harpak, A. Tradeoffs in modeling context dependency in complex trait genetics. *bioRxiv* 2023.06.21.545998 (2023) doi:10.1101/2023.06.21.545998.

111. Lambert, S. A., Abraham, G. & Inouye, M. Towards clinical utility of polygenic risk scores. *Hum. Mol. Genet.* **28**, R133–R142 (2019).

112. Ding, Y. *et al.* Polygenic scoring accuracy varies across the genetic ancestry continuum in all human populations. *bioRxiv* 2022.09.28.509988 (2022) doi:10.1101/2022.09.28.509988.

113. Johnson, R. *et al.* Leveraging genomic diversity for discovery in an electronic health record linked biobank: the UCLA ATLAS Community Health Initiative. *Genome Med.* **14**, 104 (2022).

114. Wiley, L. K. *et al.* Building a vertically-integrated genomic learning health system: The Colorado Center for Personalized Medicine Biobank. *bioRxiv* (2022) doi:10.1101/2022.06.09.22276222.

115. Belbin, G. M. *et al.* Toward a fine-scale population health monitoring system. *Cell* **184**, 2068-2083.e11 (2021).

116. Abul-Husn, N. S. & Kenny, E. E. Personalized medicine and the power of electronic health records. *Cell* **177**, 58–69 (2019).

117. Wand, H. *et al.* Improving reporting standards for polygenic scores in risk prediction studies. *Nature* **591**, 211–219 (2021).

118. Wei, J. *et al.* Calibration of polygenic risk scores is required prior to clinical implementation: results of three common cancers in UKB. *J. Med. Genet.* **59**, 243–247 (2022).

119. van Houwelingen, H. C. Validation, calibration, revision and combination of prognostic survival models. *Stat. Med.* **19**, 3401–3415 (2000).

120. Van Calster, B. *et al.* Calibration: the Achilles heel of predictive analytics. *BMC Med.* **17**, 230 (2019).

121. Sun, J. *et al.* Translating polygenic risk scores for clinical use by estimating the confidence bounds of risk prediction. *Nat. Commun.* **12**, 5276 (2021).

122. Smyth, G. K. Generalized linear models with varying dispersion. *J. R. Stat. Soc.* **51**, 47–60 (1989).

123. Koenker, R. *Quantile Regression*. (Cambridge University Press, 2005).

124. Rigby, R. A. & Stasinopoulos, D. M. Generalized additive models for location, scale and shape. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **54**, 507–554 (2005).

125. Romano, Y., Patterson, E. & Candès, E. J. Conformalized Quantile Regression. *arXiv [stat.ME]* (2019).

126. Gneiting, T. & Katzfuss, M. Probabilistic forecasting. *Annu. Rev. Stat. Appl.* **1**, 125–151 (2014).

127. Yang, J. *et al.* FTO genotype is associated with phenotypic variability of body mass index. *Nature* **490**, 267–272 (2012).

128. Young, A. I., Wauthier, F. L. & Donnelly, P. Identifying loci affecting trait variability and detecting interactions in genome-wide association studies. *Nat. Genet.* **50**, 1608–1614 (2018).

129. Miao, J. *et al.* A quantile integral linear model to quantify genetic effects on phenotypic variability. *Proc. Natl. Acad. Sci. U. S. A.* **119**, e2212959119 (2022).

130. Schoeler, T. *et al.* Participation bias in the UK Biobank distorts genetic associations and downstream analyses. *Nat. Hum. Behav.* (2023) doi:10.1038/s41562-023-01579-9.

131. Selzam, S. *et al.* Comparing within- and between-family polygenic score prediction. *Am. J. Hum. Genet.* **105**, 351–363 (2019).

132. Okbay, A. *et al.* Polygenic prediction of educational attainment within and between families from genome-wide association analyses in 3 million individuals. *Nat. Genet.* **54**, 437–449 (2022).

133. Yengo, L. *et al.* A saturated map of common genetic variants associated with human height. *Nature* **610**, 704–712 (2022).

134. Graham, S. E. *et al.* The power of genetic diversity in genome-wide association studies of lipids. *Nature* **600**, 675–679 (2021).

135. Lambert, S. A. *et al.* The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation. *Nat. Genet.* **53**, 420–425 (2021).

136. Durvasula, A. & Price, A. L. Distinct explanations underlie gene-environment interactions in the UK Biobank. *medRxiv* (2023) doi:10.1101/2023.09.22.23295969.

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

138. Patel, A. P. *et al.* A multi-ancestry polygenic risk score improves risk prediction for coronary artery disease. *Nat. Med.* **29**, 1793–1803 (2023).

139. Schumacher, F. R. *et al.* Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. *Nat. Genet.* **50**, 928–936 (2018).

140. Zhang, H. *et al.* Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses. *Nat. Genet.* **52**, 572–581 (2020).

141. Rosenberg, N. A. *et al.* Genome-wide association studies in diverse populations. *Nat. Rev. Genet.* **11**, 356–366 (2010).