Single locus theory of admixture is not enough for the study of complex traits in admixed populations

Hanbin Lee[1,2,*], Moo-Hyuk Lee[1*], Kangcheng Hou[3], Bogdan Pasaniuc[3,4,5,6] and Buhm Han[1,7,8]

[1]Department of Medicine, Seoul National University College of Medicine, Seoul

[2]Department of Mathematical Sciences, Seoul National University, Seoul

[3]Bioinformatics Interdepartmental Program, UCLA, Los Angeles

[4]Department of Human Genetics, David Geffen School of Medicine at UCLA, Los Angeles

[5]Department of Computational Medicine, David Geffen School of Medicine at UCLA, Los Angeles

[6]Department of Pathology and Laboratory Medicine, David Geffen School of Medicine at UCLA, Los Angeles

[7]Department of Biomedical Sciences, Seoul National University, Seoul

[8]Interdisciplinary Program of Bioengineering, Seoul National University, Seoul

* These authors contributed equally

Corresponding authors:

Hanbin Lee (hanbin973@snu.ac.kr) and Buhm Han (buhm.han@snu.ac.kr)

Abstract

Admixed populations offer valuable insights into the genetic architecture of complex traits. Investigation of such a topic demands a delicate theory that can handle the complex demographic history of admixed populations. With additional assumptions, we show that Pritchard-Stephens-Donnelly (PSD) can predict empirical GWAS findings. We name this the extended PSD (ePSD) model and evaluate the predictions using real data. When applied to real data of admixed genomes, the prediction of the ePSD model is very successful in explaining single-locus behaviors while falling short in predicting two-locus phenomena involving linkage disequilibrium. Our results show that a mosaic of independent single-continental segments is an insufficient approximation of contemporary admixed populations. A more advanced theory that better models linkage disequilibrium of admixed populations will be crucial to better understanding the genetic architecture of complex traits.

Introduction

The Pritchard-Stephens-Donnelly (PSD) model has successfully inferred the population structure of admixed populations[1-6]. In this model, population structure is a latent variable called the ancestral proportion (AP), also called global ancestry. Allele frequencies are then represented as weighted averages of ancestry-specific allele frequencies, where the ancestral proportion is the weights. When inferring the ancestral proportion, it is assumed that the loci used in the analysis are approximately independent without linkage disequilibrium (LD). Since the allele frequency of ancestral populations is unavailable, contemporary genomes of single continental origins are used as surrogates.

While the PSD model infers the genome-wide AP of individuals, local ancestry (LA) inference methods predict the source population of chromosome segments[7-10]. These methods view admixed genomes as a mosaic of single continental genomes. Contemporary genomes with ancestry labels are used as surrogates of the ancestral genome to identify the source population of a segment in admixed genomes.

Both global ancestry and local ancestry are used extensively in the study of complex traits of admixed populations. Global ancestry adjustment is an essential ingredient of GWAS to control for population structure[11,12]. Although principal components (PC) are generally used instead, it has been shown that PCs are merely a linear transformation of ancestral proportions, making the two adjustments equivalent[13,14]. Often, to handle the fine-scale structure of admixed genomes, LA is further included in the analysis[15-18].

The PSD model is not designed for GWAS per se. This is because GWAS relies on LD which is a two-loci property while the PSD model describes the marginal distribution of a single locus[1,2,19]. Nevertheless, a simple extension to incorporate the two-loci scenario is found in the literature. The extension assumes that the length of the local ancestry segment is far longer than the range of within-continental LD[18,20-23]. This makes the local LD structure of the segment to be the same as the source population of single continental origin. Technically speaking, it means that variants on different local ancestry segments are independent conditional on global ancestry. Examples include the TRACTOR GWAS pipeline for ancestry-specific effect estimates and several polygenic score methods tailored for admixed populations[18,20-23]. We will call the PSD model equipped with this additional assumption as the extended PSD (ePSD) model.

In this work, we show that the PSD and the ePSD model can make empirical predictions on GWAS. The standard error of GWAS methods for admixed populations can be derived from the PSD model alone. The prediction is concordant with the recent empirical findings demonstrating the high power of the conventional GWAS (the Armitage Trend Test, ATT). Furthermore, it turns out that the ancestry-specific estimates of TRACTOR are mutually independent, allowing meta-analysis of independent cohorts to be applied to these estimates to improve power that circumvents the high degrees-of-freedom of the method[24-26]. Furthermore, ePSD can make predictions about the marginal effect sizes obtained from GWAS methods. As argued by the authors of TRACTOR through simulations, TRACTOR effect sizes are the same as the effect sizes that would have been obtained had the GWAS been performed on ancestral populations separately, provided the ePSD is correct. This means that TRACTOR estimates can be directly supplied to summary statistics-based downstream analyses without further modification.

We verify the predictions of the models by applying them to admixed genomes of All of Us (AoU) and Population Architecture using Genomics and Epidemiology (PAGE). We find that

standard error predictions of the PSD model are extremely accurate with remarkably high concordance with real data ($R^2 > 0.99$). The proposed meta-analysis approach for combining TRACTOR estimates is also more powerful than the original TRACTOR test. However, the predictions of the ePSD model on effect sizes are found to be poor. We applied LDSC to summary statistics of 19 quantitative traits and measured the genetic correlation between European effect sizes produced from TRACTOR and standard GWAS of European participants of the UK Biobank. All genetic correlations were significantly smaller than 1, and LDSC even occasionally failed to run due to the negative heritability estimates from the TRACTOR estimates.

Our results show that despite the success of the PSD model in single locus problems like standard error prediction, its simple extension assuming homogeneous LD patterns limited within local ancestry segments fails to produce reliable predictions. As LD is a crucial component of the genome, the results highlight the importance of more realistic linkage disequilibrium modeling in admixed populations for understanding the genetic architecture of complex traits.

Results

The Pritchard-Stephens-Donelly model and its extension

The genotype of an individual is determined by a two-step process according to the PSD model. For each locus, *local ancestry* (LA) is first assigned according to the *global ancestry* (GA). The global ancestry is a vector of length that is equal to the number of ancestral populations summing up to 1. Each entry of the global ancestry is the probability in which a randomly selected locus in the genome has originated from a particular ancestry. Technically, it means that the distribution of the LA at a locus follows a multinomial distribution with probability equal to the GA.

Next, the genotype of the locus is assigned according to the allele frequency of the LA of the locus. Once the local ancestry is fixed, the genotypes of the two haplotypes of an individual are assumed to be independent. Therefore, the genotype follows a binomial distribution with two trials and the success probability set to the ancestry-specific allele frequency.

It is important to note that the model only describes the marginal distribution of a single locus, so the joint distribution is ignored. As genome-wide association study (GWAS) relies on linkage disequilibrium (LD), a property of the joint distribution of two loci, the original PSD model is insufficient to deal with GWAS. Hence, a simple extension has been implicitly and explicitly used in literature. We call this model the *extended PSD* (ePSD) model.

The ePSD model states that LA segments extend much further than LD within continental groups. This assumption is supported by the fact that admixture events occurred less than a few dozens of generations ago. Based on the assumption, one can assume that the local LD patterns around a particular locus is determined by the LA of the locus. As we expect markers to grab signals from adjacent causal variants, the intensity of the LD between markers and variants is then supposed to be equal to the LD of the ancestry of the LA segment. In the following sections, we investigate the consequences of both models.

The power of various admixed GWAS methods can be predicted by the PSD model

Distinct GWAS methods for admixed GWAS use different regression equations, so the effect estimates are generally not the same. Also, there is no clear agreement on the exact form of the parameters that are being estimated through these methods. Therefore, power, a function of both effect size and variance (or equivalently, the standard error), cannot be easily compared. Nevertheless, if we limit our scope to the variance only, a straightforward comparison can be derived from the PSD model.

It is a well-known fact that the variance of the regression estimator is inversely proportional to the variance of the marker, but the formula becomes complicated in general with covariates. Fortunately, a simple analytic expression is deduced for global ancestry adjustment under the PSD model (see **Methods**). The variance of the TRACTOR estimate is slightly more complicated because it involves multiple ancestry-specific allele counts of the markers in a single regression. Under the PSD model, however, we show that their variances are inversely proportional to the ancestry-specific marker variances similar to standard GWAS applied to non-admixed genomes.

By comparing the theoretical predictions to the estimated standard error, we found that the predictions are highly concordant with real data (**Figure 1a**). Furthermore, the covariances between ancestry-specific markers are exactly zero, allowing us to meta-analyze them as if they are from independent cohorts (see **Methods**). This method showed improved power

over the original Tractor statistics across various quantitative traits (**Figure 1b**).

ePSD fails to capture the pattern of real data

The authors of TRACTOR have previously shown that TRACTOR produces correct ancestry-specific marginal effect sizes through simulations. We provide mathematical proof that this observation is correct (see **Methods**). To be specific, the ancestry-specific coefficients of the TRACTOR regression equation are a product of ancestry-specific (variance-normalized) LD covariance and the effect size of the underlying causal variant. Furthermore, the theory offers the interpretation of local ancestry coefficients that were previously believed to capture confounding effects. Surprisingly, local ancestry coefficients also capture the signal from the underlying causal variant and not from confounding. Dropping the marker variables from the regression gives the interpretation of admixture mapping coefficients, uncovering the mathematical theory of admixture mapping as a byproduct.

We then attempted to verify the ePSD model through its prediction by comparing the ancestry-specific effect sizes of TRACTOR with standard GWAS estimates from African and European ancestry. The standard GWAS summary statistics were obtained from the Pan UK Biobank (PanUKBB). Then, we applied LDSC to estimate the genome-wide genetic correlation of AoU and PanUKBB summary statistics (see **Methods**). Unlike the predictions of the original PSD model, the ePSD predictions fail to explain the pattern of real data (**Figure 1c**). For 15 quantitative traits, the majority of genetic correlations are smaller than 1. Furthermore, LDSC even fails to run for some traits like Hb1Ac and Height due to negative heritability outputs.

Discussion

In this work, we mathematically derived identifiable predictions of the PSD model and its extension and then verified them using real data. These predictions illuminate the properties of a variety of GWAS methods applied to admixed genomes and suggest simple but effective improvements. Especially, the theory explains why standard GWAS regression is more powerful than methods that adjust local ancestry in an attempt to control for fine-scale population structure. On the contrary, predictions of the ePSD model turn out to be inaccurate, showing low agreement with real data.

There have been several studies comparing the power of various GWAS methods applied to admixed cohorts[27,28]. The standard GWAS regression, often called the Armitage Trend Test (ATT) has been found to be the most powerful across settings. Our mathematical result reconfirms this finding by showing that the variance of the estimator is enlarged when adjusting local ancestry like TRACTOR. We then showed that the relative power loss of TRACTOR can be ameliorated by combining ancestry-specific estimates through meta-analysis because those estimates can be shown to be independent under the PSD model. These findings are found to be highly concordant with real data.

Our study points out the problem of assuming that LA segments extend beyond the range of within-continent LD. The assumption greatly simplifies the problem by conferring independence between variants on different LA segments conditional on LA. In GWAS, it allows the isolation of the ancestry-specific marginal effect size by confining the LD within the LA segment[23]. In polygenic risk prediction, variants can be assigned ancestry-specific weights based solely on their segment ancestry[20,22]. This modeling strategy is further supported by our mathematical result. Nevertheless, it fails to stand empirical scrutiny, as shown by the low genetic correlation with single-continental GWAS.

Then why did the assumption seem to be successful in previous studies? It is likely that the overly simplistic simulation design has been causing the problem. For example, the ePSD model is part of the simulation in the original TRACTOR paper[18]. Relatively more accurate simulation algorithms based on the classic coalescent still fail because they cannot reproduce long-range LD that goes beyond LA segments[29]. Such a pattern is not likely to occur in simple admixture models assuming a single admixture event at the time of the founding of the population. Nevertheless, recent and ongoing migration events are likely to happen in real populations that can cause long-range LD[29,30].

It is worth noting that the failure of the ePSD model does not entirely discourage the use of the methods that implicitly assume the model. For example, for risk prediction purposes, the practical utility and performance of a method are not disqualified by its imperfect modeling assumption[20,22]. We only raise caution on drawing scientific conclusions on a genome-wide scale based on the ePSD model.

There are several caveats in our analysis. Firstly, the wide confidence intervals of genetic correlation analysis leave a large uncertainty in the analysis. The wide intervals stem from the small effective sample size of the European portion of admixed genomes and the African participants of the PanUKBB. Therefore, a larger cohort of admixed and African participants is required to address the issue fully. Secondly, the low genetic correlation between admixed African GWAS and standard African GWAS may come from their true difference. African genomes exhibit a substantially higher diversity level than other continental genomes[31-33]. Hence, the true genetic difference between the genomes may have caused the low genetic correlation.

Methods

Theory of admixture GWAS under the PSD model

We defer the mathematical details to the **Supplementary Note**.

Cohort description

We analyzed summary statistics of three datasets of African-European admixed individuals. PAGE study included 17,299 individuals self-identified as African American and ~6.9 million variants UK Biobank study included 4,327 individuals with African-European admixture determined by the estimated admixture proportion and ~6.6 million variants. All of Us study included 31,375 individuals with African-European admixed ancestries determined by estimated admixture proportion and with ~0.65 million variants. Detailed steps of quality control and processing can be found in Hou et al[21].

Summary statistics from the Pan UK Biobank

Summary statistics of 15 traits of UK Biobank participants of African and European ancestry were downloaded from the Pan UK Biobank repository (https://pan.ukbb.broadinstitute.org/). As the method only contained

Meta analysis

Fixed-effects and random-effects (RE2) meta-analysis were performed using the RE2C software[25,26] (https://github.com/cuelee/RE2C).

Heritability and genetic correlation estimation

Heritability and genetic correlation estimation was performed using the linkage disequilibrium score regression (LDSC) software (https://github.com/bulik/ldsc, version 1.0.1). We adjusted the sample size according to the standard error formula (18) in the **Supplementary Note**.

Acknowledgements

This research was supported by a grant of the MD-PhD/Medical Scientist Training Program through the Korea Health Industry Development (KHDI), funded by the Ministry of Health and Welfare, Republic of Korea. We thank Doc Edge (University of Southern California, Los Angeles) for providing helpful advice after reading an early version of the manuscript. We also thank Carl Veller and Graham Coop (University of California, Davis) for later discussions. This work was approved by the UK Biobank under the application ID 95237.

Figure legends

Figure 1 Predictions of the PSD and the ePSD model evaluated in real data. **a.** Comparison of predicted and estimated standard error of regression coefficients. Top panel is for height and the lower panel is for body-mass index (BMI). **b.** Quantile-Quantile (QQ) plot GWAS results of height and BMI in the PAGE cohort. **c.** The genetic correlation of 15 traits and their 95% confidence intervals estimated by LDSC. The error bar of high-density cholesterol (HDL) was truncated due to the overtly wide standard error.

Reference

1 Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-959, doi:10.1093/genetics/155.2.945 (2000).

2 Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19**, 1655-1664, doi:10.1101/gr.094052.109 (2009).

3 Gopalan, P., Hao, W., Blei, D. M. & Storey, J. D. Scaling probabilistic models of genetic variation to millions of humans. *Nat Genet* **48**, 1587-1590, doi:10.1038/ng.3710 (2016).

4 Cheng, J. Y., Mailund, T. & Nielsen, R. Fast admixture analysis and population tree estimation for SNP and NGS data. *Bioinformatics* **33**, 2148-2155, doi:10.1093/bioinformatics/btx098 (2017).

5 Cabreros, I. & Storey, J. D. A Likelihood-Free Estimator of Population Structure Bridging Admixture Models and Principal Components Analysis. *Genetics* **212**, 1009-1029, doi:10.1534/genetics.119.302159 (2019).

6 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, doi:10.1016/j.ajhg.2022.02.015 (2022).

7 Price, A. L. *et al.* Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet* **5**, e1000519, doi:10.1371/journal.pgen.1000519 (2009).

8 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, doi:10.1016/j.ajhg.2013.06.020 (2013).

9 Salter-Townshend, M. & Myers, S. Fine-Scale Inference of Ancestry Segments Without Prior Knowledge of Admixing Groups. *Genetics* **212**, 869-889, doi:10.1534/genetics.119.302139 (2019).

10 Browning, S. R., Waples, R. K. & Browning, B. L. Fast, accurate local ancestry inference with FLARE. *Am J Hum Genet* **110**, 326-335, doi:10.1016/j.ajhg.2022.12.010 (2023).

11 Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet* **2**, e190, doi:10.1371/journal.pgen.0020190 (2006).

12 Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904-909, doi:10.1038/ng1847 (2006).

13 McVean, G. A genealogical interpretation of principal components analysis. *PLoS Genet* **5**, e1000686, doi:10.1371/journal.pgen.1000686 (2009).

14 Zheng, X. & Weir, B. S. Eigenanalysis of SNP data with an identity by descent interpretation. *Theor Popul Biol* **107**, 65-76, doi:10.1016/j.tpb.2015.09.004 (2016).

15 Wang, X. *et al.* Adjustment for local ancestry in genetic association analysis of admixed populations. *Bioinformatics* **27**, 670-677, doi:10.1093/bioinformatics/btq709 (2011).

16 Zhang, J. & Stram, D. O. The role of local ancestry adjustment in association studies using admixed populations. *Genet Epidemiol* **38**, 502-515, doi:10.1002/gepi.21835 (2014).

17 Gay, N. R. *et al.* Impact of admixture and ancestry on eQTL analysis and GWAS

colocalization in GTEx. *Genome Biol* **21**, 233, doi:10.1186/s13059-020-02113-0 (2020).

18 Atkinson, E. G. *et al.* Tractor uses local ancestry to enable the inclusion of admixed individuals in GWAS and to boost power. *Nat Genet* **53**, 195-204, doi:10.1038/s41588-020-00766-y (2021).

19 Rosenberg, N. A. & Nordborg, M. A general population-genetic model for the production by population structure of spurious genotype-phenotype associations in discrete, admixed or spatially distributed populations. *Genetics* **173**, 1665-1678, doi:10.1534/genetics.105.055335 (2006).

20 Marnetto, D. *et al.* Ancestry deconvolution and partial polygenic score can improve susceptibility predictions in recently admixed individuals. *Nat Commun* **11**, 1628, doi:10.1038/s41467-020-15464-w (2020).

21 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, doi:10.1038/s41588-023-01338-6 (2023).

22 Sun, Q. *et al.* Improving polygenic risk prediction in admixed populations by explicitly modeling ancestral-specific effects via GAUDI. *bioRxiv*, doi:10.1101/2022.10.06.511219 (2022).

23 Hu, S. *et al.* Leveraging fine-scale population structure reveals conservation in genetic effect sizes between human populations across a range of human phenotypes. *bioRxiv*, doi:10.1101/2023.08.08.552281 (2023).

24 Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190-2191, doi:10.1093/bioinformatics/btq340 (2010).

25 Han, B. & Eskin, E. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *Am J Hum Genet* **88**, 586-598, doi:10.1016/j.ajhg.2011.04.014 (2011).

26 Lee, C. H., Eskin, E. & Han, B. Increasing the power of meta-analysis of genome-wide association studies to detect heterogeneous effects. *Bioinformatics* **33**, i379-i388, doi:10.1093/bioinformatics/btx242 (2017).

27 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, doi:10.1371/journal.pgen.1001371 (2011).

28 Hou, K., Bhattacharya, A., Mester, R., Burch, K. S. & Pasaniuc, B. On powerful GWAS in admixed populations. *Nat Genet* **53**, 1631-1633, doi:10.1038/s41588-021-00953-5 (2021).

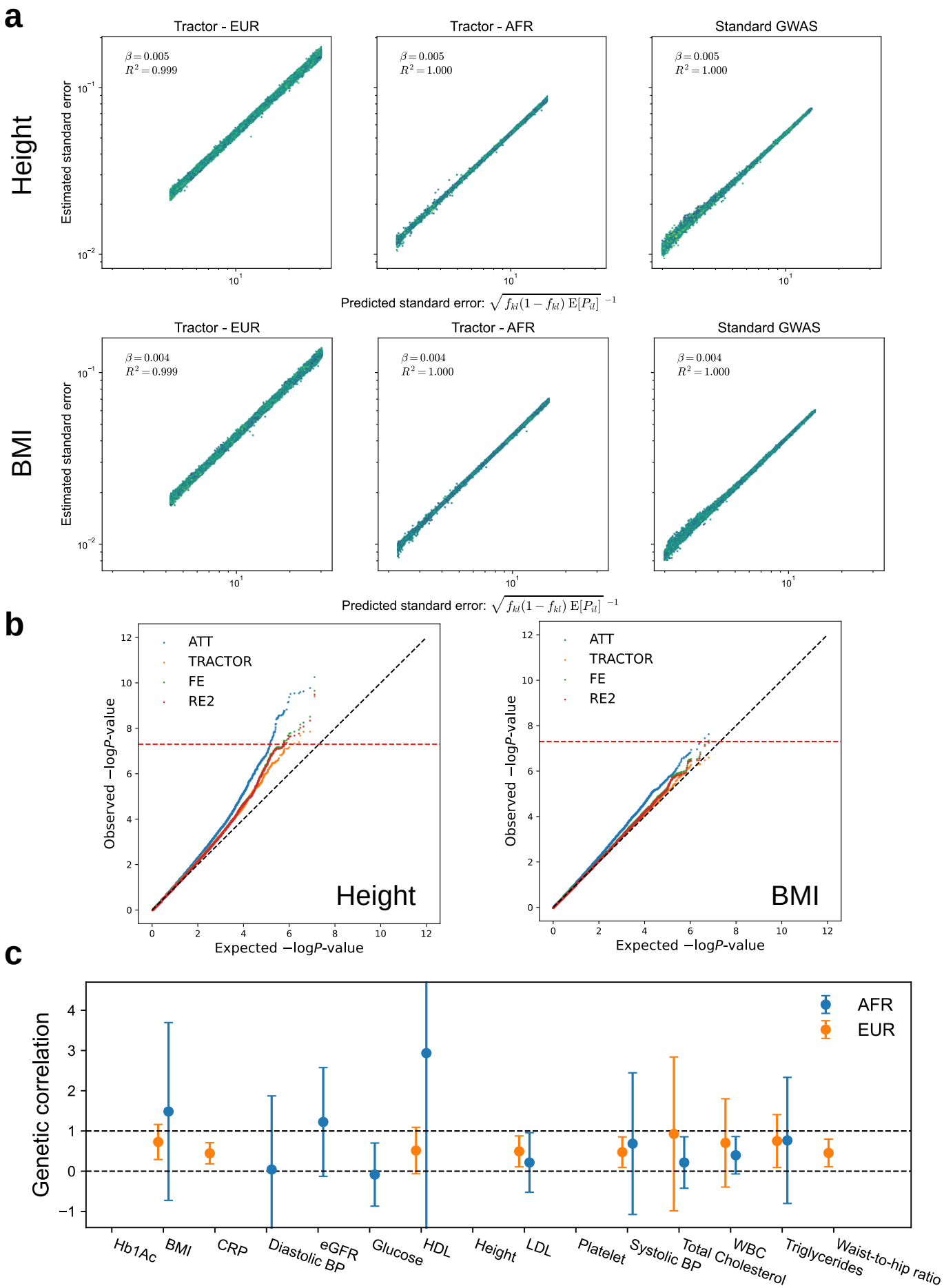
29 Nelson, D. *et al.* Accounting for long-range correlations in genome-wide simulations of large cohorts. *PLoS Genet* **16**, e1008619, doi:10.1371/journal.pgen.1008619 (2020).

30 Liang, M. & Nielsen, R. The lengths of admixture tracts. *Genetics* **197**, 953-967, doi:10.1534/genetics.114.162362 (2014).

31 Tucci, S. & Akey, J. M. The long walk to African genomics. *Genome Biol* **20**, 130, doi:10.1186/s13059-019-1740-1 (2019).

32 Fan, S. *et al.* Whole-genome sequencing reveals a complex African population demographic history and signatures of local adaptation. *Cell* **186**, 923-939 e914, doi:10.1016/j.cell.2023.01.042 (2023).

33 Gomez, F., Hirbo, J. & Tishkoff, S. A. Genetic variation and adaptation in Africa: implications for human evolution and disease. *Cold Spring Harb Perspect Biol* **6**, a008524, doi:10.1101/cshperspect.a008524 (2014).


a
Height
Tractor - EUR
Tractor - AFR
Standard GWAS
$\beta = 0.005$
$R^2 = 0.999$
$\beta = 0.005$
$R^2 = 1.000$
$\beta = 0.005$
$R^2 = 1.000$
Estimated standard error
Predicted standard error: $\sqrt{f_{kl}(1-f_{kl})\,\mathrm{E}[P_{il}]}^{\,-1}$
BMI
Tractor - EUR
Tractor - AFR
Standard GWAS
$\beta = 0.004$
$R^2 = 0.999$
$\beta = 0.004$
$R^2 = 1.000$
$\beta = 0.004$
$R^2 = 1.000$
Estimated standard error
Predicted standard error: $\sqrt{f_{kl}(1-f_{kl})\,\mathrm{E}[P_{il}]}^{\,-1}$
b
ATT
TRACTOR
FE
RE2
Observed −log*P*-value
Expected −log*P*-value
Height
ATT
TRACTOR
FE
RE2
Observed −log*P*-value
Expected −log*P*-value
BMI
c
Genetic correlation
AFR
EUR
Hb1Ac
BMI
CRP
Diastolic BP
eGFR
Glucose
HDL
Height
LDL
Platelet
Systolic BP
Total Cholesterol
WBC
Triglycerides
Waist-to-hip ratio