Estimating heritability and its enrichment in tissue-specific gene sets in admixed populations

Yang Luo[1-5,*], Xinyi Li[1-5,*], Xin Wang[6], Steven Gazal[3,7], Josep Maria Mercader[3,8], 23andMe Research Team[6], SIGMA Type 2 Diabetes Consortium, Benjamin M. Neale[3,9], Jose C. Florez[3,8,10], Adam Auton[6], Alkes L. Price[3,7,11], Hilary K. Finucane[3,#], Soumya Raychaudhuri[1-5,12,#]

Affiliations

[1]Division of Rheumatology, Immunology, and Allergy, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA
[2]Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA
[3]Broad Institute of MIT and Harvard, Cambridge, MA, USA
[4]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA
[5]Center for Data Sciences, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA
[6]23andMe, Inc., Mountain View, California, USA
[7]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA
[8]Diabetes Unit and Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA 02114, USA
[9]Analytic and Translational Genetics Unit, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA
[10]Department of Medicine, Harvard Medical School, Boston, MA, USA
[11]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA
[12]Arthritis Research UK Centre for Genetics and Genomics, Manchester Academic Health Science Centre, University of Manchester, Manchester, UK

*: These authors contributed equally to this work.

[#]: Correspondence should be addressed to H.K.F. (finucane@broadinstitute.org) or S.R. (soumya@broadinstitute.org).

Abstract

The increasing size and diversity of genome-wide association studies provide an exciting opportunity to study how the genetics of complex traits vary between diverse populations. Here, we introduce covariate-adjusted LD score regression (cov-LDSC), a method to accurately estimate genetic heritability (h_g^2) and its enrichment in both homogenous and admixed populations with summary statistics and in-sample LD estimates. In-sample LD can be estimated from a subset of the GWAS samples, allowing our method to be applied efficiently to very large cohorts. In simulations, we show that unadjusted LDSC underestimates h_g^2 by 10%- 60% in admixed populations; in contrast, cov-LDSC is robust to all simulation parameters. We apply cov-LDSC to genotyping data from approximately 170,000 Latino, 47,000 African American and 135,000 European individuals. We estimate h_g^2 and detect heritability enrichment in three quantitative and five dichotomous phenotypes respectively, making this, to our knowledge, the most comprehensive heritability-based analysis of admixed individuals. Our results show that most traits have high concordance of h_g^2 and consistent tissue-specific heritability enrichment between different ethnic groups. For example, estimates of h_g^2 for BMI are 0.22 ± 0.01, 0.23 ± 0.03 and 0.22 ± 0.01 in Latino, African American and European populations respectively. However, for age at menarche, we observe population-specific heritability differences with estimates of h_g^2. We observe consistent patterns of tissue-specific heritability enrichment across populations, for example in the limbic system for BMI, the per-standardized-annotation effect size τ^* are 0.16 ± 0.04, 0.28 ± 0.11 and 0.18 ± 0.03 in Latino, African American and European populations respectively. Our results demonstrate

that our approach is a powerful way to analyze genetic data for complex traits from underrepresented populations.

Introduction

It is important for human geneticists to study how genetic variants that influence phenotypic variability act across different populations worldwide[1,2]. With increasingly large and diverse genetic studies, it is now becoming feasible to assess how the genetic mechanisms of complex traits act across populations. However, to date, most genome-wide association studies (GWAS) have been focused on relatively homogenous continental populations and, in particular those of European descent[3]. Non-European populations, particularly those with mixed ancestral backgrounds such as African Americans and Latinos, have been underrepresented in genetic studies. Many statistical methods to analyze genetic data assume homogeneous populations. In order to ensure that the benefits of GWAS are shared beyond individuals of homogeneous continental ancestry, statistical methods for admixed populations are needed[4].

Among methods to analyze polygenic complex traits in homogeneous populations, summary statistics-based methods such as linkage disequilibrium score regression (LDSC)[5,6] and its extensions[5–8] have become particularly popular due to their computational efficiency, relative ease of application, and their applicability without raw genotyping data[9]. These methods can be used to estimate SNP heritability, the proportion of phenotypic variance explained by genotyped variants[5,10–12], distinguish polygenicity from confounding[5], establish relationships between complex phenotypes[7] , and model genome-wide polygenic signals to identify key cell types and regulatory mechanisms of human diseases[6,13,14].

Summary statistics-based methods for polygenic analysis frequently rely on accurate linkage disequilibrium (LD) calculations, which can be obtained on a random subset of the individuals without phenotypic information. Consequently, LDSC, SumHer[15] and other summary statistics-based methods have not easily been expanded to admixed populations. The LD score for a SNP is the sum of its pairwise correlations (r^2) with all other SNPs. For most homogenous populations, LD scores are easy to calculate for SNPs in local windows since r^2 between SNPs more than one centimorgan (cM) is rarely nonzero. In admixed populations however, LD scores are confounded since pairwise r^2 between SNPs is inflated due to admixture itself. Furthermore, in homogenous populations LD scores can be easily calculated in reference panels since representative data is widely available. For admixed populations, even when reference panels are available, they may not be representative of the precise populations used in the genetic study. For example Latino populations in different regions worldwide can be the result of admixture between the same continental populations, but with dramatic differences in admixture proportions and timing of the admixture event[16]. A generic reference panel cannot easily capture these differences and hence cannot produce accurate LD scores that can be widely used for all Latino populations. Thus, LDSC has only been recommended to be applied in homogeneous populations.

In this work, we first examine the performance of LDSC in admixed populations and demonstrate that LDSC yields severely downwardly biased estimates of SNP heritability. Next, we extend the LDSC-based methods to admixed populations by introducing covariate-adjusted LDSC (cov-LDSC). In cov-LDSC, for each variant we regress the global genotypic PCs, obtained within the GWAS samples, out of the raw genotypes to obtain covariate-adjusted genotypes. We then compute LD scores on the adjusted genotypes and use LDSC to estimate

heritability. Using covariate-adjusted in-sample LD to compute LD scores removes the potential concerns of reference panel mismatch, long-distance admixture-LD, and covariate effects listed above, and produces accurate estimates of heritability with summary statistics (**Methods, Figure 1**). Furthermore, heritability can be partitioned to identify key gene sets that have disproportionately high heritability. While access to the genotype data of the GWAS samples is required to compute the covariate-adjusted LD scores, LD can be estimated on a random subset of the individuals, preserving the computational efficiency of LDSC and allowing for its application to very large studies. Individual cohorts can also release the in-sample covariate-adjusted LD scores as well as the summary statistics to avoid privacy concerns associated with genotype-level information to facilitate future studies.

After demonstrating that cov-LDSC is robust to a wide range of simulation scenarios, we apply it to 8,124 Latinos from a type 2 diabetes study (the Slim Initiative in Genomic Medicine for the Americas, SIGMA)[17] as well as 161,894 Latino, 46,844 African American, and 134,999 European research participants from a personal genetics company (23andMe). We analyze three quantitative phenotypes (body mass index, height, and age at menarche), and five dichotomous phenotypes (type 2 diabetes (available in the SIGMA cohort only), left handedness, morning person, motion sickness, and nearsightedness).

One powerful component of LDSC is that it can be used to test whether a particular genome annotation -- for example, sets of genes that are specifically expressed within a candidate tissue or cell type -- capture more heritability than expected by chance[12,13]. We demonstrate that cov-LDSC can be applied in the same way to identify trait-relevant tissue and cell types in admixed and homogenous populations. We examine height, BMI and morning person since

these traits had sufficient statistical power for cell-type enrichment analyses in the personal genetics cohort. We observe a high level of consistency among enriched tissue types, highlighting that the underlying biological processes are shared among studied populations. This analysis of hundreds of genome annotations in cohorts of over 100,000 individuals would have been impossible with existing genotype-based methods[18,19].

Material and Methods

Mathematical framework of cov-LDSC

In the standard polygenic model on which LDSC is based, genotype vectors are represented as $x_1, ..., x_N \in \mathbb{R}^M$, with each SNP normalized to mean zero and variance one in the population. In the absence of covariates, we model the phenotypes y_i

$$y_i = x_i \cdot \beta + \epsilon_i \tag{1}$$

where $\epsilon_1, ..., \epsilon_N \overset{iid}{\sim} \mathcal{N}(0, \sigma_e^2)$ and $\beta \in \mathbb{R}^M$ is a vector of per-normalized-genotype effect sizes, which we model as random with mean zero. In standard LDSC, the variance of β_j is $h_g{}^2/M$; in stratified LD score regression the variance of β_j depends on a set of genome annotations. We assume, without loss of generality, that y_i has mean zero and variance one in the population.

Let $\chi_j{}^2$ denote the chi-square statistic for the j-th SNP, approximately equal to $(X_j^T Y)^2/N$, where $X_j = (x_{1j}, ..., x_{Nj})^T$ and $Y = (y_1, ..., y_n)^T$. The main equation on which LDSC is based is:

$$\mathbb{E}[\chi_j^2] \approx 1 + Na + \frac{h_g^2}{M}\ell(j). \tag{2}$$

Where the LD score, $\ell(j)$, is:

$$\ell(j) = \sum R_{jk}^2,$$

R_{jk}^2 is the correlation between SNPs j and k in the underlying population. We estimate h_g^2 via weighted regression of χ_j^2 on our estimates of $\ell(j)$, evaluating significance with a block jackknife across SNPs[6].

In the absence of covariates, the LD scores can be estimated from an external reference panel such as 1000 Genomes, as long as the correlation structure in the reference panel matches the correlation structure of the sample. In most homogeneous populations, we can also assume that the true underlying correlation is negligible outside of a 1-cM window.

Next, suppose that we would like to include covariates in our model, such that Equation (1) is replaced with

$$y_i = x_i \cdot \beta + c_i \cdot \eta + \epsilon_i \tag{3}$$

or in matrix form,

$$Y = X\beta + C\eta + \epsilon_i \tag{4}$$

where c_i is the vector of covariate values for individual i, η is a vector of effect sizes of the covariates, X is the genotype matrix with $X_{ij} = x_{ij}$, and C is a matrix of covariates with $C_{ij} = c_{ij}$. We can project the covariates out of this equation by multiplying $P = I - C(C^TC)^{-1}C^T$ on the left to get

$$\tilde{Y} = \tilde{X}\beta + \tilde{\epsilon}, \tag{5}$$

where $\tilde{Y} = PY$, $\tilde{X} = PX$ and $\tilde{\epsilon} = P\epsilon$ (if the covariates are genotype principal components, then $P = 1 - CC^T$). When sample size is large relative to the number of covariates, then the $\tilde{\epsilon}$ are still approximately independent and normally distributed, and the variance of $\tilde{y}_i$ is still

approximately one, so we can proceed as before, replacing our phenotype vector Y with $\tilde{Y}$ and our genotype matrix X with $\tilde{X}$.

The summary statistics in this covariate-adjusted model are the standard summary statistics of a GWAS with covariates, and the relevant LD is the correlation structure of $\tilde{X}$, not X. If X is a homogeneous population, then the correlation structure of $\tilde{X}$ will be similar to the correlation structure of X, and well-approximated by a reference panel. However if X is the genotype matrix from an admixed population and the covariates include PCs, then the correlation structure of $\tilde{X}$ is no longer well-approximated either by X or by a reference panel. Thus, in cov-LDSC, we compute LD scores directly from $\tilde{X}$ or a random subsample thereof. We call them the covariate-adjusted LD scores.

Using genotype data to compute LD scores means that the model being fit is based on the joint effects of a sparser set of SNPs, e.g. the genotyped SNPs, than when sequence data is used to compute LD scores. For estimating total SNP heritability, this means that cov-LDSC estimates the same estimand as GCTA (${h_g}^2$) and not the usual estimand of LDSC (${h_{common}}^2$; see below). For partitioned heritability, the density of reference panel SNPs can be important because the joint effect of a SNP in an annotation can include the tagged effect of an untyped SNP that is not in the annotation, deflating estimates of enrichment. Thus, we recommend using cov-LDSC only on annotations made of large contiguous regions, such as gene sets. Moreover, we urge caution when interpreting quantitative estimates of heritability enrichment. Here, we look at the significance of the conditional enrichment (i.e., regression coefficient) of gene sets for our tissue-specific analysis (see below).

Window size and number of PCs in LD score calculations

In addition to computing LD from the covariate-adjusted genotypes, we also investigate the appropriate window size for estimating LD scores. To do this, we examine the effect of varying the genomic window size for both simulated and real data sets. We determine that LD score estimates were robust to the choice of window size if the increase in the mean LD score estimates was less than 1% per cM beyond a given window. Using this criterion, we use window sizes of 5-cM and 20-cM for the simulated and real genotypes respectively (**Supplementary Table 1-3**). We also calculate the squared correlations between LD score estimates using the chosen window size and other LD score estimates with window sizes larger than the chosen window. The squared correlations were greater than 0.99 in all cases (**Supplementary Table 4-6**) indicating the LD score estimates were robust at the chosen window sizes.

Similarly, to determine the number of PCs needed to be included in the GWAS association tests and cov-LDSC calculations, we examine the effect of varying the genomic window size using different numbers of PCs. The number of PCs that needed to be included for covariate adjustment depended on the population structure for different datasets.

Genotype simulations

We evaluate the performance of LDSC and cov-LDSC with simulated phenotypes and both simulated and real genotypes. For the simulated genotypes, we used msprime[20] version 0.6.1 to simulate population structure with mutation rate 2×10^{-8} and recombination maps from the HapMap Project[21]. We adapt the demographic model from Mexican migration history[22] using parameters that were previously inferred from the 1000 Genomes Project[23]. We assume the

205 admixture event happened approximately 500 years ago to mirror the European colonization of
206 the Americas. We set different admixture proportions to reflect different admixed populations. In
207 each population, we simulate 10,000 individuals after removing second degree related samples
208 (kinship>0.125) using KING[24].

Slim Initiative in Genomic Medicine for the Americas (SIGMA) Type 2 Diabetes (T2D) cohort

209 8,214 Mexican and other Latin American samples were genotyped with Illumina HumanOmni2.5
210 array. We further filter the genotyped data to be MAF >5% and remove SNPs in high LD regions
211 (**Supplementary Table 7**). After QC, a total of 8,214 individuals and 943,244 SNPs remain. We
212 estimate the in-sample LD score with a 20-cM window and 10 PCs in all scenarios.
213
214 We use these genotypes for simulations. We also analyze three phenotypes from the SIGMA
215 cohort: height, BMI, and type 2 diabetes (T2D). For T2D, we assume a reported prevalence in
216 Mexico of 0.144[17]. For each phenotype, we include age, sex, and the first 10 PCs as fixed
217 effects in the association analyses.

Phenotype simulations

218 We simulate phenotypes with two different polygenic genetic architectures, given by GCTA[19]
219 and the baseline model[6] respectively. In the GCTA model, all variants are equally likely to be
220 causal independent of their functional or minor allele frequency (MAF) structure, and the
221 standardized causal effect size variance is constant, i.e. $var(\beta_j) = h_g^2/M$. In contrast, the
222 baseline model incorporates functionally dependent architectures. Briefly, it includes 53
223 overlapping genome-wide functional annotations (e.g. coding, conserved, regulatory). It models

$var(\beta_j) = \sum_C \alpha_c(j)\tau_c$ where $\alpha_c(j)$ is the value of annotation α_c at variant j and τ_c represents the per-variant contribution, of one unit of the annotation α_c, to heritability. We generate all causal variants among common observed variants with MAF >5% (~40,000 SNPs in simulated genotypes and 943,244 SNPs in the SIGMA cohort). To represent environmental stratification, similar to previously described[5], we add 0.2 * standardized first principal component to the standardized phenotypes.

We simulate both quantitative and case-control traits with both GCTA and baseline model genetic architectures, using both simulated and real genotypes, varying the number of causal variants, the true heritability, and environmental stratification. For case-control simulations, we adopt a liability threshold model with disease prevalence 0.1. We obtain 5,000 cases and 5,000 controls for each simulation scenario.

To obtain summary statistics for the simulated traits, we apply single-variant linear models for quantitative traits and logistic models for binary trait both with 10 PCs as covariates in association analyses using PLINK 1.90[25].

23andMe cohort

All participants were drawn from the customer base of 23andMe, Inc., a direct to consumer genetics company. Participants provided informed consent and participated in the research online, under a protocol approved by the external AAHRPP-accredited IRB, Ethical & Independent Review Services (www.eandireview.com). Samples from 23andMe are then chosen from consented individuals who were genotyped successfully on an Illumina Infinium Global Screening Array (~640,000 SNPs) supplemented with ~50,000 SNPs of custom content.

We restrict participants to those who have European, African American, or Latino ancestry determined through an analysis of local ancestry[26].

To compute LD scores, we use both genotyped and imputed SNPs. We filter genotyped variants with a genotype call rate $\leq$ 90%, non-zero self-chain score, strong evidence of Hardy Weinberg disequilibrium ($p > 10^{-20}$), and failing a parent-offspring transmission test. For imputed variants, we use a reference panel that combined the May 2015 release of the 1000 Genomes Phase 3 haplotypes[23] with the UK10K imputation reference panel[27]. Imputed dosages are rounded to the nearest integer (0, 1, 2) for downstream analysis. We filter variants with imputation r-squared $\leq$ 0.9. We also filter genotyped and imputed variants for batch effects and sex dependent effects. To minimize rounding inaccuracies, we prioritize genotyped SNPs over imputed SNPs in the merged SNP set. We restrict the merged SNP set to HapMap3 variants with MAF $\geq$ 0.05. We measure LD scores in a subset of African Americans (61,021) and Latinos (9,990) on chromosome 2 with different window sizes from 1-cM to 50-cM (**Supplementary Table 3**) and squared correlation between different window sizes (**Supplementary Table 6)**. We compute all LD scores with a 20-cM window.

In genome-wide association analyses, for each population, we choose a maximal set of unrelated individuals for each analysis using a segmental identity-by-descent (IBD) estimation algorithm[28]. We define individuals to be related if they share more than 700-cM IBD.

We performe association tests using linear regression model for quantitative traits and logistic regression model for binary traits assuming additive allelic effects. We include covariates for

age, sex and the top 10 PCs to account for residual population structure. We list details of phenotypes and genotypes in **Supplementary Table 8.**

Heritability Estimation

We calculate in-sample LD scores using both a non-stratified LD score model and the baseline model[6]. In simulated phenotypes generated with the GCTA model, we use non-stratified LDSC to estimate heritability. In simulated phenotypes generated using the baseline model, we use LDSC-baseline to estimate heritability. We use the 53 non-frequency dependent annotations included in the baseline model to estimate h_g^2 in the 23andMe research database and the SIGMA cohort real phenotypes. We recognize recent studies have shown that genetic heritability can be sensitive to the choice of LD-dependent heritability model[8,11]. However, understanding the LD- and MAF-dependence of complex trait genetic architecture is an important but complex endeavor potentially requiring both modeling of local ancestry as well as large sequenced reference panels that are currently unavailable. We thus leave this complexity for future work.

h_g^2 versus h_{common}^2

The quantity (h_g^2) we reported in the main analysis is defined as heritability tagged by HapMap3 variants with MAF $\geq$5%, including tagged causal effects of both low-frequency and common variants. This quantity is different from h_{common}^2, the heritability casually explained by all common SNPs excluding tagged causal effects of low-frequency variants, reported in the original LDSC[5]. In Europeans and other homogeneous populations, it is possible to estimate h_{common}^2, since reference panels, such as 1000 Genome Project, are available which include >99% of the SNPs with frequency >1%[23]. However, in-sample sequence data is usually not

available for an admixed GWAS cohort, and so cov-LDSC can only include genotyped SNPs in the reference panel, and thus can only estimate the heritability tagged by a given set of genotyped SNPs. In order to compare the same quantity across cohorts, we use common HapMap3 SNPs (MAF $\geq$ 5%) for in-sample LD reference panel calculation, since most of them should be well imputed for a genome-wide genotyping array. To quantify the difference between ${h_g}^2$ and ${h_{common}}^2$, we pre-phase the genotype data in the SIGMA cohort using SHAPEIT2[29]. We use IMPUTE2[30] to impute genotypes at untyped genetic variants using the 1000 Genomes Project Phase 3[23] dataset as a reference panel. We merge genotyped SNPs and all well imputed (INFO>0.99) SNPs (~6.9 million) in the SIGMA cohort as a reference panel and reported ${h_{common}}^2$, to approximate what the estimate of ${h_{common}}^2$ would have been with a sequenced reference panel (**Supplementary Table 9**).

Tissue Type Specific Analyses

Following Finucane et al[13], we extend cov-LDSC so that we can assess enrichment in and around sets of genes that are specifically expressed in tissue and cell-types (cov-LDSC-SEG). We annotate the genes with the same set of tissue specific expressed genes identified previously[13] using the Genotype–Tissue Expression (GTEx) project and a public dataset made available by the Franke lab. We calculate within-sample stratified cov-LD scores in the 23andMe cohort for each of these gene sets. We obtain regression coefficients $\hat{\tau}_c$ from the model and normalize them as

$$\tau_c^* = \frac{M_{h_g^2} \cdot sd_c}{h_g^2} \hat{\tau}_c,$$

Where ${M_{h_g}}^2$ $M_{h_g^2}$ is the number of SNPs used to calculate ${h_g}^2$ and sd_c is the standard deviation (sd) of annotation a_c[8]. We interpret τ_c^* as the proportional change of averaged

310 per-SNP heritability by one sd increase in value of the annotation of each cell type, conditional
311 on other 53 non-cell type specific baseline annotations. We calculate a one-tailed p-value for
312 each coefficient where the null hypothesis is that the coefficient is non-positive[13]. All the
313 significant enrichments are reported with false discovery rate < 5% $(-log_{10}(p) > 2.75)$. We
314 perform fixed-effect inverse variance weighting meta-analysis using τ_c^* and normalized standard
315 error (se) across populations.

Software Availability

316 An open-source software implementation of covariate-adjusted LD score regression is publicly
317 available (see Web Resources).

Results

Robustness of LD score estimation

318 To demonstrate the effect of admixture on the stability of LD score estimates, we first calculated
319 LD scores with genomic window sizes ranging from 0-50 cM in both European (EUR, N=503)
320 and admixed American (AMR, N=347) populations from the 1000 Genomes Project[23]. As
321 window size increases, we expect the mean LD score to plateau because LD should be
322 negligible for large enough distance. If the mean LD score does not plateau, but continues to
323 rise with increasing window size, then one of two possibilities may apply: (1) the window is too
324 small to capture all of the LD; (2) the LD scores are capturing long-range pairwise SNP
325 correlations arising from admixture. If this increase is non-linear then there is non-negligible
326 distance-dependent LD, violating LDSC assumptions. Examining unadjusted LD scores, we
327 observed that in the EUR population[5], the mean LD score estimates plateaued at windows

beyond 1-cM in size, as previously reported. However, in the AMR population the mean LD score estimates continued to increase concavely with increasing window size. In contrast, when we applied cov-LDSC with 10 PCs to calculate covariate adjusted LD scores, we observed that LD score estimates plateaued for both EUR and AMR at a 1-cM and 20-cM window size respectively (<1% increase per cM, **Supplementary Table 10**). This suggested that cov-LDSC was able to correct the long-range LD due to admixture and yielded stable estimates of LD scores (**Method, Supplementary Figure 1**), and also that cov-LDSC was applicable in homogeneous populations (**Supplementary Table 10**). The larger window size for the AMR population was needed due to residual LD caused by recent admixture. We next tested the sensitivity of the LD score estimates with regard to the number of PCs included in the cov-LDSC. We observed that in the AMR panel, LD score estimates were unaffected by adding PCs and by increasing window sizes above 20-cM (**Supplementary Figure 2**).

Simulations with simulated genotypes

To assess whether cov-LDSC produces unbiased estimates of h_g^2, we first simulated genotypes of admixed individuals (**Methods**). We simulated genotypes of 10,000 unrelated diploid individuals for approximately 400,000 common SNPs on chromosome 2 in a coalescent framework using msprime[20]. First, we tested LDSC and cov-LDSC with different admixture proportions between two ancestral populations, and a quantitative phenotype with a h_g^2 of 0.4 using an additive model (**Methods**). We observed that as the proportion of admixture increased, $\hat{h}_g^2$ for LDSC increasingly underestimated true h_g^2 by as much as 18.6%. In marked contrast, cov-LDSC produced consistently unbiased estimates regardless of admixture proportion (**Supplementary Figure 3a**).

Second, we varied the percentage of causal variants from 0.01% to 50% in a polygenic quantitative trait with $h_g^2 = 0.4$ in a population with a fixed admixture proportion of 50%. LDSC again consistently underestimated h_g^2 by 12%-18.6%. In contrast, cov-LDSC yielded unbiased estimates regardless of the percentage of causal variants (**Supplementary Figure 3b**).

Third, we assessed the robustness of LDSC and cov-LDSC for different assumed total h_g^2 (0.05, 0.1, 0.2, 0.3, 0.4 and 0.5). At each h_g^2 value, LDSC underestimated by 11.5%-19.6%. For cov-LDSC, we observed that the standard error increased with h_g^2, but point estimates remained unbiased (**Supplementary Figure 3c**).

Fourth, we included an environmental stratification component aligned with the first PC of the genotype data (**Methods**), and concluded that cov-LDSC was also robust to confounding (**Supplementary Figure 3d**).

Finally, to assess the performance of cov-LDSC in polygenic binary phenotypes, we simulated genotype data for a binary trait with a prevalence of 0.1 assuming a liability threshold model (**Methods**). We showed that cov-LDSC provided unbiased estimates in case-control studies with the same four simulation scenarios (**Supplementary Figure 4**). In contrast, LDSC underestimated heritability for binary phenotypes in the same way as it did for quantitative phenotypes.

Simulation results with real genotypes

We next examined the performance of both unadjusted LDSC and cov-LDSC on real genotypes of individuals from admixed populations. We used genotype data from SIGMA, which includes

372 8,214 Mexican and other Latino individuals. Using ADMIXTURE[31] and populations from the
373 1000 Genomes Project as reference panels, we observed that each individual in the SIGMA
374 cohort had different admixture proportions (**Supplementary Figure 5**). As in the AMR panel, we
375 observed that using a 20-cM window, LD score estimates plateaued in the SIGMA cohort
376 (**Supplementary Figure 6, Supplementary Table 2**), and were unaffected by different numbers
377 of PCs (**Supplementary Figure 7**). When we assumed a non-infinitesimal, additive model with
378 1% of all SNPs to be causal and h_g^2 =0.4, we observed that cov-LDSC h_g^2 estimates produced
379 unbiased estimates using a 20-cM window with 10 PCs (**Supplementary Figure 8**). We
380 subsequently used a 20-cM window and 10 PCs in all simulations.
381
382 We observed that cov-LDSC yielded unbiased h_g^2 estimates in simulated traits where we varied
383 the number of causal variants and total heritabilities (**Figure 2a-b**). In contrast, LDSC
384 underestimated heritability by as much as 62.5%. To examine the performance of cov-LDSC in
385 the presence of environmental confounding factors, we simulated an environmental stratification
386 component aligned with the first PC of the genotype data, representing European v.s. Native
387 American ancestry. In this simulation scenario, cov-LDSC still provided unbiased h_g^2 estimates
388 (**Figure 2c**). Intercepts of all the simulation scenarios were close to 1, suggesting that we had
389 adequately controlled for confounding from population stratification and cryptic relatedness
390 (**Supplementary Figure 9a-c**). We observed unbiased h_g^2 estimates with both the GCTA model
391 and baseline model (**Supplementary Table 11**).
392
393 Thus far, we have used cov-LDSC by calculating LD scores on the same set of samples that
394 were used for association studies (in-sample LD scores). In practical applications, computing LD
395 scores on the whole data set can be computationally expensive and difficult to obtain, so we

396 investigated computing LD scores on a subset of samples. To investigate the minimum number
397 of samples required to obtain accurate in-sample LD scores, we computed LD scores on
398 subsamples of 100, 500, 1,000 and 5,000 individuals from a GWAS of 10,000 simulated
399 genotypes (**Supplementary Figure 10**). We repeated these analyses in simulated phenotypes
400 in the SIGMA cohort. We subsampled the SIGMA cohort, and obtained unbiased estimates
401 when using as few as 1,000 samples (**Figure 2d**). We therefore recommend computing
402 in-sample LD scores on a randomly chosen subset of at least 1,000 individuals from a GWAS in
403 our approach.

Application to SIGMA and 23andMe cohorts

404 We next used cov-LDSC to estimate h_g^2 of height, BMI and T2D phenotypes, measured within
405 the SIGMA cohort (**Methods, Table 1**). We estimated h_g^2 of height, BMI and T2D to be 0.38 ±
406 0.08, 0.25 ± 0.06 and 0.26 ± 0.08 respectively. These results were similar to reported values
407 from UK Biobank[32] and other studies[10,33] for European populations. Although estimands differed
408 in different studies (**Methods**), we noted that without cov-LDSC, we would have obtained
409 severely deflated estimates (**Table 1**). To confirm that our reported heritability estimates were
410 robust under different model assumptions, we applied an alternative approach based on REML
411 in the linear mixed model framework implemented in GCTA[19]. To avoid biases introduced from
412 calculating genetic relatedness matrices (GRMs) in admixed individuals, we obtained a GRM
413 based on an admixture-aware relatedness estimation method REAP[34] (**Methods**). GCTA-based
414 results were similar to reported h_g^2 estimates from cov-LDSC, indicating our method was able
415 to provide reliable h_g^2 estimates in admixed populations (**Table 1**). We noted, however, that the
416 GCTA-based results would be impossible to obtain on the much larger datasets, for example the
417 23andMe cohort described below.

We then applied both LDSC and cov-LDSC to 161,894 Latino, 46,844 African American and 134,999 European research participants from 23andMe. We analyzed three quantitative and four dichotomous phenotypes (**Methods**). In this setting, summary statistic methods to estimate heritability were essential since the dataset would be too large for computationally expensive genotype-based strategies. We used a 20-cM window and 10 PCs in LD score calculations for both populations (**Supplementary Figure 11**). LDSC and cov-LDSC produced similar heritability estimates in the European population, whereas in the admixed populations, LDSC consistently provided low estimates of h_g^2. Heritability estimates were similar between non-stratified and baseline models (**Supplementary Table 12**). For each phenotype, we estimated h_g^2 using the same population-specific in-sample LD scores. For most phenotypes, the reported h_g^2 was similar among the three population groups with a notable exception for age at menarche (**Figure 3**). This suggested possible differences ($p = 7.1 \times 10^{-3}$ between Latinos and Europeans) in the genetic architecture of these traits between different ethnic groups. It has been long established that there is population variation in the timing of menarche[35,36]. Early menarche might influence the genetic basis of other medically relevant traits since early age at menarche is associated with a variety of chronic diseases such as childhood obesity, coronary heart disease and breast cancer[37,38]. These results highlighted the importance of including diverse populations in genetic studies in order to enhance our understanding of complex traits that show differences in their genetic heritability.

Tissue type specific analysis

We applied stratified cov-LDSC to sets of specifically expressed genes[13] (SEG) to identify trait-relevant tissue and cell types in traits included in the 23andMe cohort across European,

440 Latino, and African American populations. We only tested height, BMI and morning person,
441 which were the three traits that had heritability z-scores larger than seven in at least two
442 populations[6] (**Supplementary Table 13**). We also performed inverse-variance weighting
443 meta-analysis across the three populations (**Supplementary Table 14**). Across different
444 populations, BMI showed consistent enrichment in central nervous system gene sets . In the
445 European population, most of the enrichments recapitulated the results from the previous
446 analysis using UK Biobank[13]. We found similar but fewer enrichments in Latinos and African
447 Americans, most likely due to smaller sample sizes. The most significantly enriched tissue types
448 for BMI in all three populations were limbic system ($\tau^*_{EUR} = 0.18$, $\tau^*_{LAT} = 0.16$, $\tau^*_{AA} = 0.28$,
449 $\tau^*_{meta} = 0.18$), entorhinal cortex ($\tau^*_{EUR} = 0.18$, $\tau^*_{LAT} = 0.15$, $\tau^*_{AA} = 0.24$, $\tau^*_{meta} = 0.17$), and
450 cerebral cortex ($\tau^*_{EUR} = 0.16$, $\tau^*_{LAT} = 0.14$, $\tau^*_{AA} = 0.15$, $\tau^*_{meta} = 0.15$); none of the three effects
451 were significantly different across populations. When we compared the enrichments for all of the
452 tissues between population pairs, we observed that they were similar with regression slopes
453 close to one (European vs. Latinos: slope=0.69, se=0.036, $p = 3.1 \times 10^{-48}$; European vs.
454 African Americans: slope=0.67, se=0.12, $p = 3.4 \times 10^{-8}$) (**Figure 4(a)-(e)**). The sizes of these
455 three brain structures have been shown to be correlated with BMI using magnetic resonance
456 imaging data[39]. The midbrain and the limbic system are highly involved in the food rewarding
457 signals through dopamine releasing pathway[40]. Furthermore, the hypothalamus in the limbic
458 system releases hormones that regulate appetite, energy homeostasis and metabolisms, like
459 leptin, insulin, and ghrelin[40,41]. For height, similar to previously reported associations[13], we also
460 identified enrichments in the gene sets derived from musculoskeletal and connective tissues. In
461 the meta-analysis, the three most significant enrichments were cartilage ($\tau^*_{EUR} = 0.21$,
462 $\tau^*_{LAT} = 0.19$, $\tau^*_{AA} = 0.24$, $\tau^*_{meta} = 0.20$), chondrocytes ($\tau^*_{EUR} = 0.21$, $\tau^*_{LAT} = 0.15$, $\tau^*_{AA} = 0.11$,

463 $\tau^*_{meta} = 0.17$), and uterus ($\tau^*_{EUR} = 0.17$, $\tau^*_{LAT} = 0.15$, $\tau^*_{AA} = 0.16$, $\tau^*_{meta} = 0.16$) (**Figure 4(f)-(j)**).
464 A heterogeneity test revealed no difference across three populations (I^2< 70% and p-value >
465 0.05). The slopes of the regression of the two population coefficients were close to 1 (European
466 vs. Latinos: slope = 0.76, se=0.020, $p = 4.9 \times 10^{-94}$; European vs. African Americans: 0.59,
467 se=0.055, $p = 1.4 \times 10^{-21}$). The importance of these tissues and their roles in height have been
468 addressed in the previous pathway analysis, expression quantitative trait loci (eQTLs) and
469 epigenetic profiling[42,43]. Previous studies have shown that the longitudinal growth of bones is
470 partly controlled by the number and proliferation rate of chondrocytes on the growth plate which
471 is a disc of cartilages[44]. For the morning person phenotype, we found enrichments in many brain
472 tissues in Europeans, concordant with a previous study[45]. Entorhinal cortex ($\tau^*_{EUR} = 0.16$,
473 $\tau^*_{LAT} = 0.22$, $\tau^*_{meta} = 0.18$), cerebral cortex ($\tau^*_{EUR} = 0.15$, $\tau^*_{LAT} = 0.22$, $\tau^*_{meta} = 0.18$), and brain (
474 $\tau^*_{EUR} = 0.17$, $\tau^*_{LAT} = 0.19$, $\tau^*_{meta} = 0.18$) were enriched in both Latinos and Europeans.
475 Evidence showed that circadian rhythm was controlled by the suprachiasmatic nucleus, the
476 master clock in our brain, and also the circadian oscillator that resides in neurons of the cerebral
477 cortex [46,47,48]. We also found unique enrichments of esophagus muscularis and the esophagus
478 gastroesophageal junction in the Latino populations, but the heterogeneity test showed that the
479 difference is not significant (I^2 = 0.49 and 0.50 respectively). We observed that the slope of
480 regression between Latino coefficients and European coefficients across gene sets was 0.70
481 (se=0.062, $p = 1.9 \times 10^{-23}$) (**Figure 4(k)-(n)**). Compared to the original LDSC-SEG,
482 cov-LDSC-SEG increased the statistical power in African Americans and Latinos
483 (**Supplementary Figure 12-14**).

Discussion

As we expand genetic studies to explore admixed populations around the world, extending statistical genetics methods to make inferences within admixed populations is crucial. This is particularly true for methods based on summary statistics, which are dependent on the use of LD scores that we showed to be problematic in admixed populations. In this study, we confirmed that original LDSC and other summary statistics-based methods, such as PCGCs[49] and SumHer[50] that were originally designed for homogenous populations, should not be applied to admixed populations. We introduced cov-LDSC which regresses out global PCs on individual genotypes during the LD score calculation, and showed it can yield unbiased LD scores, heritability estimates and its enrichment, such as trait-relevant cell and tissue type enrichments, in homogenous and admixed populations.

Although our work provides a novel, efficient approach to estimate genetic heritability and to identify trait-relevant cell and tissue types using summary statistics in admixed populations, it has a few limitations (**Methods**). First, covariates included in the summary statistics should match the covariates included in the covariate-adjusted LD score calculations (**Supplementary Figure 15**), and h_g^2 estimates in admixed populations are more sensitive to potentially unmatched LD reference panels. Unmatched reference panels are likely to produce biased estimates[51,52]. We examined the performance of using an out-of-sample reference panel in admixed populations (See **Appendix**) and caution that when using 1000 Genomes or any out-of-sample reference panels for a specific admixed cohort, users should ensure that the demographic histories are shared between the reference and the study cohort. Large sequencing projects such as TOPMed[53] that include large numbers (N>1,000) of admixed

samples can potentially serve as out-of-sample LD reference panels, although further investigations are needed to study their properties. We therefore advise to compute in-sample LD scores from the full or a random subset of data (N>1,000) used to generate the admixed GWAS summary statistics when possible. We also strongly encourage cohorts to release their summary statistics and in-sample covariate-adjusted LD scores at the same time to facilitate future studies. Second, when applying cov-LDSC to imputed variants, particularly those with lower imputation accuracy (INFO <0.99), we caution that the heritability estimates can be influenced by an imperfect imputation reference panel, especially in Latino populations[54,55]. To limit the bias in varying genotyping array and imputation quality in studied admixed cohorts, we recommend restricting the heritability analyses to common HapMap3 variants. Any extension to a larger set of genetic variants, especially across different cohorts should be performed with caution. Third, when we evaluated the performance of cov-LDSC in case-control studies, we assumed no presence of binary covariates with strong effects and demonstrated that cov-LDSC can yield robust h_g^2 estimates. However, it has been shown that LDSC can provide biased estimates in the presence of extreme ascertainment for dichotomous phenotypes[49]. Adapting cov-LDSC into case-control studies under strong binary effects remains a potential avenue for future work. Fourth, recent studies have shown that heritability estimates can be sensitive to the choice of the LD- and frequency-dependent heritability model[8,11,12,15]. Since our approach can flexibly add annotations to estimate heritability under the model that is best supported by the data, we believe it provides a good foundation for addressing the question of how to incorporate ancestry-dependent frequencies in the LD-dependent annotation in the future (**Methods**).

Despite these limitations, in comparison with other methods, such as those based on restricted maximum likelihood estimation (REML)[19] with an admixture-aware GRM, for estimating h_g^2 in

admixed populations or those with intra-population structure, cov-LDSC has a number of attractive properties. First, covariate-adjusted in-sample LD scores can be obtained with a subset of samples, enabling analysis of much larger cohorts than was previously possible. Second, LD scores only need to be calculated once per cohort; this is particularly useful in large cohorts such as 23andMe and UK Biobank[56], where multiple phenotypes have been collected per individual and per-trait heritability can be estimated based on the same LD scores. Second, as a generalized form of LDSC, it is robust to population stratification and cryptic relatedness in both homogenous and admixed populations. Third, similar to the original LDSC methods, cov-LDSC can be extended to perform analyses such as estimating genetic correlations, partitioning h_g^2 by functional annotations, identifying disease-relevant tissue and cell types and multi-trait analysis[6,57,58].

By applying cov-LDSC to approximately 344,000 individuals from European, African American, and Latin American ancestry, we observed evidence of heritability differences across different populations. These differences highlight the importance of studying diverse populations. How these differences may correspond to differences in biological mechanisms may lead to mechanistic insights about the phenotype. One strategy to do this, which we explored by extending cov-LDSC, is to partition heritability by different cell type- and tissue-specific annotations to dissect the genetic architecture in admixed populations. Our results demonstrated that although there are some cases of nominal heterogeneity across populations among tested tissue-types, most of the tissue-specific enrichments are consistent among the populations studied here. This is consistent with the previous findings that show strong correspondence in functional and cell type enrichment between Europeans and Asians[59,60]. Seeing the same tissue-type for a single trait emerge in multiple populations can give us more

confidence that this tissue may account for polygenic heritability. Larger sample sizes are needed to increase the power of our current analyses and to enhance our understanding of how genetic variants that are responsible for heritable phenotypic variability differ among populations.

As the number of admixed and other diverse GWAS and biobank data become readily available[1,53,61], our approach provides a powerful way to study admixed populations.

Appendix

In-sample versus out-of-sample LD

To test the reliability of using an out-of-sample reference LD panel for cov-LDSC applications, we first examined the performance of out-of-sample LD scores obtained from 1,000 samples with a perfectly matching demographic history in the simulated genotypes. cov-LDSC yielded unbiased estimates when using 1,000 samples in an out-of-sample reference panel with a perfectly matching population structure (**Supplementary Figure 10**). Next, we tested the accuracy of heritability estimates when using 1000 Genome Project to obtain out-of-sample LD scores. When using the AMR panel as a reference panel for the SIGMA cohort, we observed an unbiased h_g^2 estimate ($p = 0.33$, **Figure 2d**). This suggested that the AMR panel included in the 1000 Genomes Project has similar demographic history compared to the SIGMA cohort (**Supplementary Figure 5**). However, as we decreased the number of samples included in the subsampling, the cov-LDSC regression intercepts deviated further from 1 (**Supplementary Figure 9d**). This is probably due to attenuation bias from noisily estimated LD scores at N<1,000. Next, we explored the feasibility of applying an out-of-sample reference panel in cell

and tissue type specific analyses. To this end, we used 1000 Genomes AMR samples (N=347) to obtain stratified LD scores and applied it on summary statistics obtained from 23andMe. In contrast to using in-sample LD scores, we discovered no significant tissue type enrichment (**Supplementary Figure 16**) when using LD scores obtained from the 1000 Genomes. This suggested an out-of-sample reference panel may significantly reduce the power for detecting trait relevant tissue types in admixed populations.

We therefore caution that when using 1000 Genomes or any out-of-sample reference panels for a specific admixed cohort, users should ensure that the demographic histories are shared between the reference and the study cohort. We highly recommend computing in-sample LD scores on a randomly chosen subset of at least 1,000 individuals from a GWAS. We also strongly encourage cohorts to release their summary statistics and in-sample covariate-adjusted LD scores at the same time to facilitate future studies.

Web Resources

cov-LDSC software and tutorials, https://github.com/immunogenomics/cov-ldsc

msprime, https://pypi.python.org/pypi/msprime;

GCTA, http://cnsgenomics.com/software/gcta/;

LDSC, https://github.com/bulik/ldsc/;

PLINK 1.90, https://www.cog-genomics.org/plink2;

REAP v1.2, http://faculty.washington.edu/tathornt/software/REAP/download.html;

ADMIXTURE v1.3.0, http://www.genetics.ucla.edu/software/admixture/download.html;

Acknowledgements

The study was supported by the National Institutes of Health (NIH) TB Research Unit Network, Grant U19 AI111224-01. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

We thank the research participants and employees of 23andMe for their contribution to this study.

Author contributions

Y.L., X.L, H.K.F and S.R. conceived and supervised the study. Y.L. and X.L. analyzed data. X.W. and A.A. contributed and analyzed the 23andMe data. S.G., B.M.N. and A.L.P. gave critical feedback on LDSC and statistical models. J.M.M. and J.C.F. contributed the SIGMA study. All authors contributed to the writing of this manuscript.

Members of the 23andMe Research Team:
Michelle Agee, Babak Alipanahi, Robert K. Bell, Katarzyna Bryc, Sarah L. Elson, Pierre Fontanillas, Nicholas A. Furlotte, Barry Hicks, David A. Hinds, Karen E. Huber, Ethan M. Jewett, Yunxuan Jiang, Aaron Kleinman, Keng-Han Lin, Nadia K. Litterman, Matthew H. McIntyre, Kimberly F. McManus, Joanna L. Mountain, Elizabeth S. Noblin, Carrie A.M. Northover, Steven J. Pitts, G. David Poznik, J. Fah Sathirapongsasuti, Janie F. Shelton, Suyash Shringarpure, Chao Tian, Joyce Y. Tung, Vladimir Vacic, and Catherine H. Wilson.

Competing Financial interests

X.W., A.A. and members of the 23andMe Research Team are employees of 23andMe, Inc., and hold stock or stock options in 23andMe.

References

1. Sirugo, G., Williams, S.M., and Tishkoff, S.A. (2019). The Missing Diversity in Human Genetic Studies. Cell *177*, 26–31.

2. Martin, A.R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B.M., and Daly, M.J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. Nat. Genet. *51*, 584–591.

3. Popejoy, A.B., and Fullerton, S.M. (2016). Genomics is failing on diversity. Nature *538*, 161–164.

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

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

6. Finucane, H.K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., Anttila, V., Xu, H., Zang, C., Farh, K., et al. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat. Genet. *47*, 1228–1235.

7. Bulik-Sullivan, B., Finucane, H.K., Anttila, V., Gusev, A., Day, F.R., Loh, P.-R., ReproGen Consortium, Psychiatric Genomics Consortium, Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium 3, Duncan, L., et al. (2015). An atlas of genetic correlations across human diseases and traits. Nat. Genet. *47*, 1236–1241.

8. Gazal, S., Finucane, H.K., Furlotte, N.A., Loh, P.-R., Palamara, P.F., Liu, X., Schoech, A., Bulik-Sullivan, B., Neale, B.M., Gusev, A., et al. (2017). Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. Nat. Genet. *49*, 1421–1427.

9. Pasaniuc, B., and Price, A.L. (2017). Dissecting the genetics of complex traits using summary association statistics. Nat. Rev. Genet. *18*, 117–127.

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

11. Speed, D., Cai, N., UCLEB Consortium, Johnson, M.R., Nejentsev, S., and Balding, D.J. (2017). Reevaluation of SNP heritability in complex human traits. Nat. Genet. *49*, 986–992.

12. Gazal, S., Marquez-Luna, C., Finucane, H.K., and Price, A.L. Reconciling S-LDSC and LDAK models and functional enrichment estimates.

647 13. Finucane, H.K., Reshef, Y.A., Anttila, V., Slowikowski, K., Gusev, A., Byrnes, A., Gazal, S.,
648 Loh, P.-R., Lareau, C., Shoresh, N., et al. (2018). Heritability enrichment of specifically
649 expressed genes identifies disease-relevant tissues and cell types. Nat. Genet. *50*, 621–629.

650 14. Guo, J., Yang, J., and Visscher, P.M. (2018). Leveraging GWAS for complex traits to detect
651 signatures of natural selection in humans. Curr. Opin. Genet. Dev. *53*, 9–14.

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

654 16. Homburger, J.R., Moreno-Estrada, A., Gignoux, C.R., Nelson, D., Sanchez, E., Ortiz-Tello,
655 P., Pons-Estel, B.A., Acevedo-Vasquez, E., Miranda, P., Langefeld, C.D., et al. (2015). Genomic
656 insights into the ancestry and demographic history of South America. PLoS Genet. *11*,
657 e1005602.

658 17. SIGMA Type 2 Diabetes Consortium, Williams, A.L., Jacobs, S.B.R., Moreno-Macías, H.,
659 Huerta-Chagoya, A., Churchhouse, C., Márquez-Luna, C., García-Ortíz, H., Gómez-Vázquez,
660 M.J., Burtt, N.P., et al. (2014). Sequence variants in SLC16A11 are a common risk factor for
661 type 2 diabetes in Mexico. Nature *506*, 97–101.

662 18. Zaitlen, N., Pasaniuc, B., Sankararaman, S., Bhatia, G., Zhang, J., Gusev, A., Young, T.,
663 Tandon, A., Pollack, S., Vilhjálmsson, B.J., et al. (2014). Leveraging population admixture to
664 characterize the heritability of complex traits. Nat. Genet. *46*, 1356–1362.

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

667 20. Kelleher, J., Etheridge, A.M., and McVean, G. (2016). Efficient Coalescent Simulation and
668 Genealogical Analysis for Large Sample Sizes. PLoS Comput. Biol. *12*, e1004842.

669 21. International HapMap Consortium (2003). The International HapMap Project. Nature *426*,
670 789–796.

671 22. Gutenkunst, R.N., Hernandez, R.D., Williamson, S.H., and Bustamante, C.D. (2009).
672 Inferring the joint demographic history of multiple populations from multidimensional SNP
673 frequency data. PLoS Genet. *5*, e1000695.

674 23. 1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P.,
675 Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., et al. (2015). A global
676 reference for human genetic variation. Nature *526*, 68–74.

677 24. Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., and Chen, W.-M. (2010).
678 Robust relationship inference in genome-wide association studies. Bioinformatics *26*,
679 2867–2873.

680 25. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015).
681 Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience *4*,
682 7.

683 26. Durand, E.Y., Do, C.B., Mountain, J.L., and Michael Macpherson, J. (2014). Ancestry

Composition: A Novel, Efficient Pipeline for Ancestry Deconvolution.

27. UK10K Consortium, Walter, K., Min, J.L., Huang, J., Crooks, L., Memari, Y., McCarthy, S., Perry, J.R.B., Xu, C., Futema, M., et al. (2015). The UK10K project identifies rare variants in health and disease. Nature *526*, 82–90.

28. Henn, B.M., Hon, L., Macpherson, J.M., Eriksson, N., Saxonov, S., Pe'er, I., and Mountain, J.L. (2012). Cryptic distant relatives are common in both isolated and cosmopolitan genetic samples. PLoS One *7*, e34267.

29. O'Connell, J., Gurdasani, D., Delaneau, O., Pirastu, N., Ulivi, S., Cocca, M., Traglia, M., Huang, J., Huffman, J.E., Rudan, I., et al. (2014). A general approach for haplotype phasing across the full spectrum of relatedness. PLoS Genet. *10*, e1004234.

30. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., and Abecasis, G.R. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat. Genet. *44*, 955–959.

31. Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Res. *19*, 1655–1664.

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

33. Xue, A., Wu, Y., Zhu, Z., Zhang, F., Kemper, K.E., Zheng, Z., Yengo, L., Lloyd-Jones, L.R., Sidorenko, J., Wu, Y., et al. (2018). Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. Nat. Commun. *9*, 2941.

34. Thornton, T., Tang, H., Hoffmann, T.J., Ochs-Balcom, H.M., Caan, B.J., and Risch, N. (2012). Estimating Kinship in Admixed Populations. Am. J. Hum. Genet. *91*, 122–138.

35. Demerath, E.W., Liu, C.-T., Franceschini, N., Chen, G., Palmer, J.R., Smith, E.N., Chen, C.T.L., Ambrosone, C.B., Arnold, A.M., Bandera, E.V., et al. (2013). Genome-wide association study of age at menarche in African-American women. Hum. Mol. Genet. *22*, 3329–3346.

36. Fernández-Rhodes, L., Malinowski, J.R., Wang, Y., Tao, R., Pankratz, N., Jeff, J.M., Yoneyama, S., Carty, C.L., Setiawan, V.W., Le Marchand, L., et al. (2018). The genetic underpinnings of variation in ages at menarche and natural menopause among women from the multi-ethnic Population Architecture using Genomics and Epidemiology (PAGE) Study: A trans-ethnic meta-analysis. PLoS One *13*, e0200486.

37. Canoy, D., Beral, V., Balkwill, A., Wright, F.L., Kroll, M.E., Reeves, G.K., Green, J., Cairns, B.J., and Million Women Study Collaborators* (2015). Age at menarche and risks of coronary heart and other vascular diseases in a large UK cohort. Circulation *131*, 237–244.

38. Bodicoat, D.H., Schoemaker, M.J., Jones, M.E., McFadden, E., Griffin, J., Ashworth, A., and Swerdlow, A.J. (2014). Timing of pubertal stages and breast cancer risk: the Breakthrough Generations Study. Breast Cancer Res. *16*, R18.

39. Taki, Y., Kinomura, S., Sato, K., Inoue, K., Goto, R., Okada, K., Uchida, S., Kawashima, R.,

721 and Fukuda, H. (2008). Relationship Between Body Mass Index and Gray Matter Volume in
722 1,428 Healthy Individuals. Obesity *16*, 119–124.

723 40. Berthoud, H.-R., Münzberg, H., and Morrison, C.D. (2017). Blaming the Brain for Obesity:
724 Integration of Hedonic and Homeostatic Mechanisms. Gastroenterology *152*, 1728–1738.

725 41. Clemmensen, C., Müller, T.D., Woods, S.C., Berthoud, H.-R., Seeley, R.J., and Tschöp, M.H.
726 (2017). Gut-Brain Cross-Talk in Metabolic Control. Cell *168*, 758–774.

727 42. Guo, M., Liu, Z., Willen, J., Shaw, C.P., Richard, D., Jagoda, E., Doxey, A.C., Hirschhorn, J.,
728 and Capellini, T.D. (2017). Epigenetic profiling of growth plate chondrocytes sheds insight into
729 regulatory genetic variation influencing height. Elife *6*,.

730 43. Wood, A.R., Esko, T., Yang, J., Vedantam, S., Pers, T.H., Gustafsson, S., Chu, A.Y.,
731 Estrada, K., Luan, J. 'an, Kutalik, Z., et al. (2014). Defining the role of common variation in the
732 genomic and biological architecture of adult human height. Nat. Genet. *46*, 1173–1186.

733 44. Villemure, I., and Stokes, I.A.F. (2009). Growth plate mechanics and mechanobiology. A
734 survey of present understanding. J. Biomech. *42*, 1793–1803.

735 45. Jones, S.E., Lane, J.M., Wood, A.R., van Hees, V.T., Tyrrell, J., Beaumont, R.N., Jeffries,
736 A.R., Dashti, H.S., Hillsdon, M., Ruth, K.S., et al. (2019). Genome-wide association analyses of
737 chronotype in 697,828 individuals provides insights into circadian rhythms. Nat. Commun. *10*,
738 343.

739 46. Bering, T., Carstensen, M.B., Wörtwein, G., Weikop, P., and Rath, M.F. (2018). The
740 Circadian Oscillator of the Cerebral Cortex: Molecular, Biochemical and Behavioral Effects of
741 Deleting the Arntl Clock Gene in Cortical Neurons. Cereb. Cortex *28*, 644–657.

742 47. Gnocchi, D., and Bruscalupi, G. (2017). Circadian Rhythms and Hormonal Homeostasis:
743 Pathophysiological Implications. Biology *6*,.

744 48. Potter, G.D.M., Skene, D.J., Arendt, J., Cade, J.E., Grant, P.J., and Hardie, L.J. (2016).
745 Circadian Rhythm and Sleep Disruption: Causes, Metabolic Consequences, and
746 Countermeasures. Endocr. Rev. *37*, 584–608.

747 49. Weissbrod, O., Flint, J., and Rosset, S. (2018). Estimating SNP-Based Heritability and
748 Genetic Correlation in Case-Control Studies Directly and with Summary Statistics. Am. J. Hum.
749 Genet. *103*, 89–99.

750 50. Speed, D., and Balding, D. (2018). Better estimation of SNP heritability from summary
751 statistics provides a new understanding of the genetic architecture of complex traits.

752 51. Yang, J., Bakshi, A., Zhu, Z., Hemani, G., Vinkhuyzen, A.A.E., Nolte, I.M., van
753 Vliet-Ostaptchouk, J.V., Snieder, H., Lifelines Cohort Study, Esko, T., et al. (2015).
754 Genome-wide genetic homogeneity between sexes and populations for human height and body
755 mass index. Hum. Mol. Genet. *24*, 7445–7449.

756 52. Yang, J., Zeng, J., Goddard, M.E., Wray, N.R., and Visscher, P.M. (2017). Concepts,
757 estimation and interpretation of SNP-based heritability. Nat. Genet. *49*, 1304–1310.

53. Taliun, D., Harris, D.N., Kessler, M.D., Carlson, J., Szpiech, Z.A., Torres, R., Gagliano Taliun, S.A., Corvelo, A., Gogarten, S.M., Kang, H.M., et al. (2019). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program.

54. Martin, A.R., Gignoux, C.R., Walters, R.K., Wojcik, G.L., Neale, B.M., Gravel, S., Daly, M.J., Bustamante, C.D., and Kenny, E.E. (2017). Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. Am. J. Hum. Genet. *100*, 635–649.

55. Nelson, S.C., Stilp, A.M., Papanicolaou, G.J., Taylor, K.D., Rotter, J.I., Thornton, T.A., and Laurie, C.C. (2016). Improved imputation accuracy in Hispanic/Latino populations with larger and more diverse reference panels: applications in the Hispanic Community Health Study/Study of Latinos (HCHS/SOL). Hum. Mol. Genet. *25*, 3245–3254.

56. Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al. (2015). UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. *12*, e1001779.

57. Turley, P., Walters, R.K., Maghzian, O., Okbay, A., Lee, J.J., Fontana, M.A., Nguyen-Viet, T.A., Wedow, R., Zacher, M., Furlotte, N.A., et al. (2018). Multi-trait analysis of genome-wide association summary statistics using MTAG. Nat. Genet.

58. Gusev, A., Lee, S.H., Trynka, G., Finucane, H., Vilhjálmsson, B.J., Xu, H., Zang, C., Ripke, S., Bulik-Sullivan, B., Stahl, E., et al. (2014). Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. Am. J. Hum. Genet. *95*, 535–552.

59. Kichaev, G., and Pasaniuc, B. (2015). Leveraging Functional-Annotation Data in Trans-ethnic Fine-Mapping Studies. Am. J. Hum. Genet. *97*, 260–271.

60. Kanai, M., Akiyama, M., Takahashi, A., Matoba, N., Momozawa, Y., Ikeda, M., Iwata, N., Ikegawa, S., Hirata, M., Matsuda, K., et al. (2018). Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. Nat. Genet. *50*, 390–400.

61. Matise, T.C., Ambite, J.L., Buyske, S., Carlson, C.S., Cole, S.A., Crawford, D.C., Haiman, C.A., Heiss, G., Kooperberg, C., Marchand, L.L., et al. (2011). The Next PAGE in understanding complex traits: design for the analysis of Population Architecture Using Genetics and Epidemiology (PAGE) Study. Am. J. Epidemiol. *174*, 849–859.

Figure and table legends

Figure 1. Overview of the covariate-adjusted LD score regression. (a) As input, cov-LDSC takes raw genotypes of collected GWAS samples and their global principal components. (b) cov-LDSC regresses out the ancestral components based on global principal components from the LD score calculation and corrects for long-range *admixture* LD. Black and red lines indicate estimates before and after covariate adjustment respectively (c) Adjusted heritability estimation based on GWAS association statistics (measured by χ^2) and covariate-adjusted LD scores. (d) Estimation of heritability enrichment in tissue-specific gene sets.

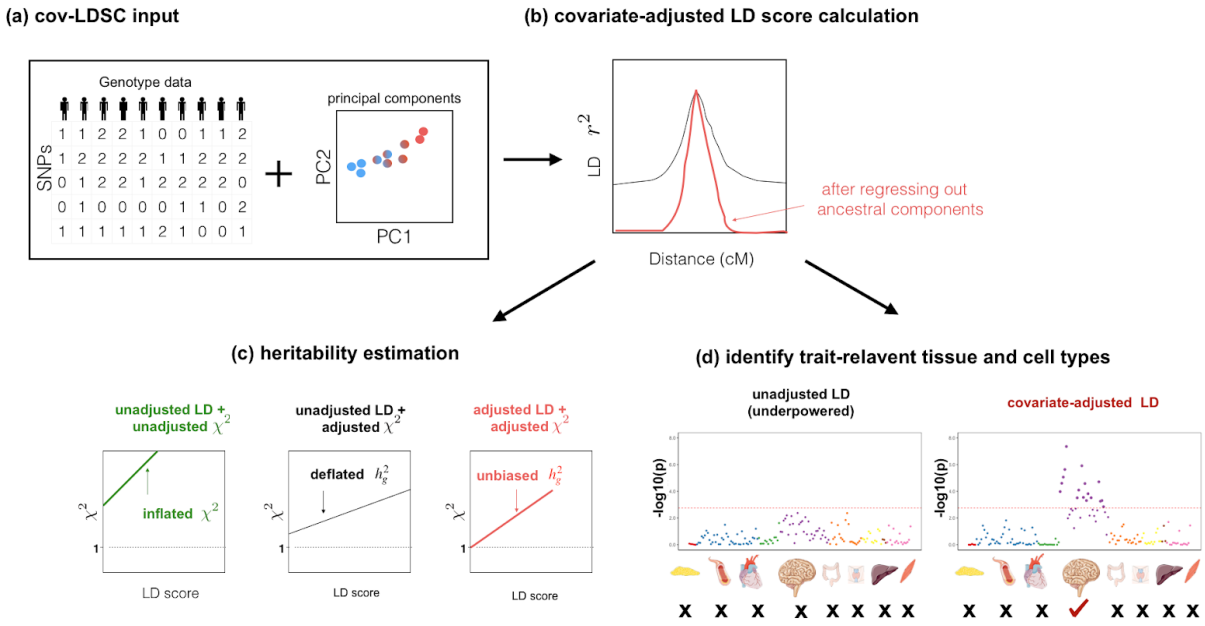

Figure 2. Estimates of heritability (h_g^2) under different simulation scenarios using the SIGMA cohort. LDSC (orange) underestimated h_g^2 and cov-LDSC (blue) yielded robust h_g^2 estimates under all settings. Each boxplot represents the mean LD score estimate from 100 simulated phenotypes using the genotypes of 8,214 unrelated individuals from the SIGMA cohort. For cov-LDSC, a window size of 20-cM with 10 PCs were used in all scenarios. A true polygenic quantitative trait with $h_g^2 = 0.4$ is assumed for scenarios (a), (c) and (d) and 1% causal variants are assumed for scenarios (b)-(d). (a) h_g^2 estimation with varying proportions of causal variants (0.01% - 30%). (b) h_g^2 estimation with varying heritabilities (0, 0.05, 0.1, 0.2, 0.3, 0.4 and 0.5). (c) h_g^2 estimation when an environmental stratification component aligned with the first PC of the genotype data was included in the phenotype simulation. (d) h_g^2 estimation when using a subset of the cohort to obtain LD score estimates and using out-of-sample LD score estimates obtained from Admixed Americans included in the 1000 Genomes Project.

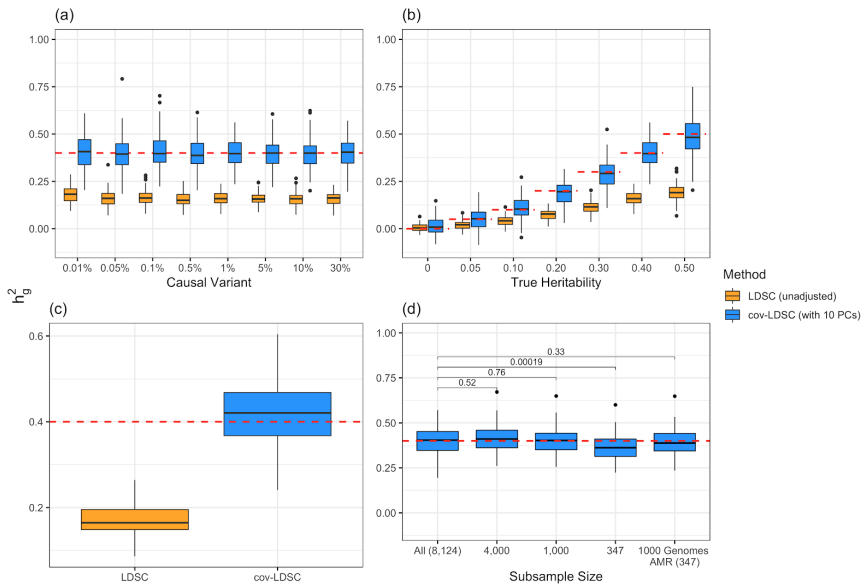

Figure 3. Estimates of heritability (h_g^2) of three quantitative and four dichotomous traits in two admixed populations in the 23andMe research cohort. For seven selected non-disease phenotypes (body mass index (BMI), height, age at menarche, left handedness, morning person, motion sickness and nearsightedness) in the 23andMe cohort, we reported their estimated genetic heritabilities and intercepts (and their standard errors) using the baseline model. LD scores were calculated using 134,999, 161,894, 46,844 individuals from 23andMe European, Latino and African American individuals respectively. For each trait, we reported the sample size in obtained summary statistics used in cov-LDSC. For BMI and height, we also reported the h_g^2 estimates from the SIGMA cohort.

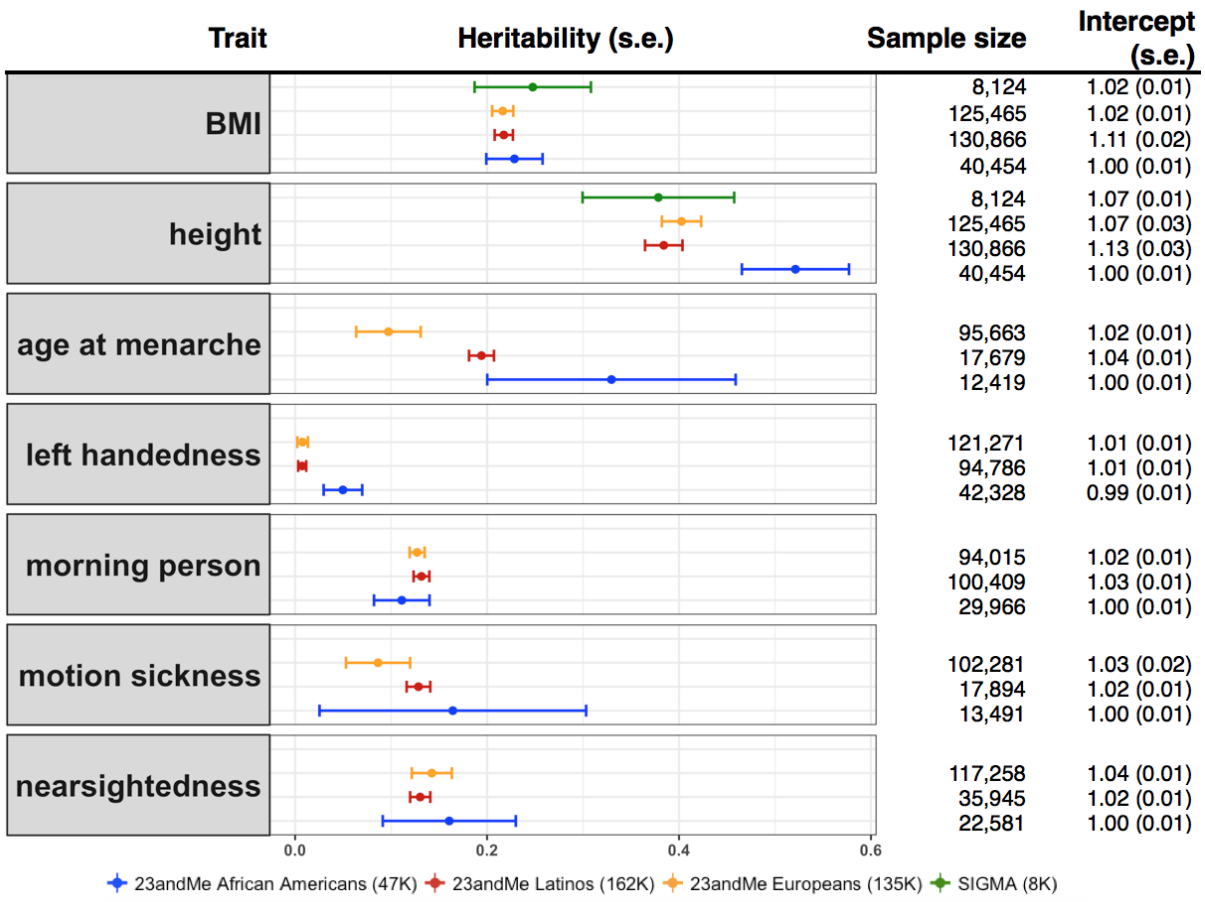

Figure 4. Results of multiple-tissue analysis for height, BMI and morning person. Each point represents a tissue type from either the GTEx data set or the Franke lab data set as defined in Finucane et al[13]. From top to bottom, (a)-(d) show multiple-tissue analysis for BMI in the cross-population meta-analysis and in Europeans, Latinos and African Americans respectively. (e) shows the scatter plot of the estimated per-standardized-annotation effect size τ^* which represents the proportional change of averaged per-SNP heritability by one standard deviation increase in value of the annotation of each cell type, conditional on other 53 non-cell type specific baseline annotations, in the three populations for all tested tissue types (**Methods**). The x-axis shows the τ^* in European populations and the y-axis shows either τ^* in Latinos (blue) or African Americans (orange). We reported the slope and p-value when we regress Latinos (blue) and African Americans (organe) τ^* on Europeans τ^* for all tissue types. Error bars indicate standard errors of τ^*. Similarly, the results are shown in (f)-(j) for height and (k)-(n) for morning person. The significance threshold in plots (a)-(d), (f-i) and (k-m) is defined by the FDR < 5% cutoff, $-log_{10}(p) = 2.75$. Numerical results are reported in **Supplementary Table 14**.

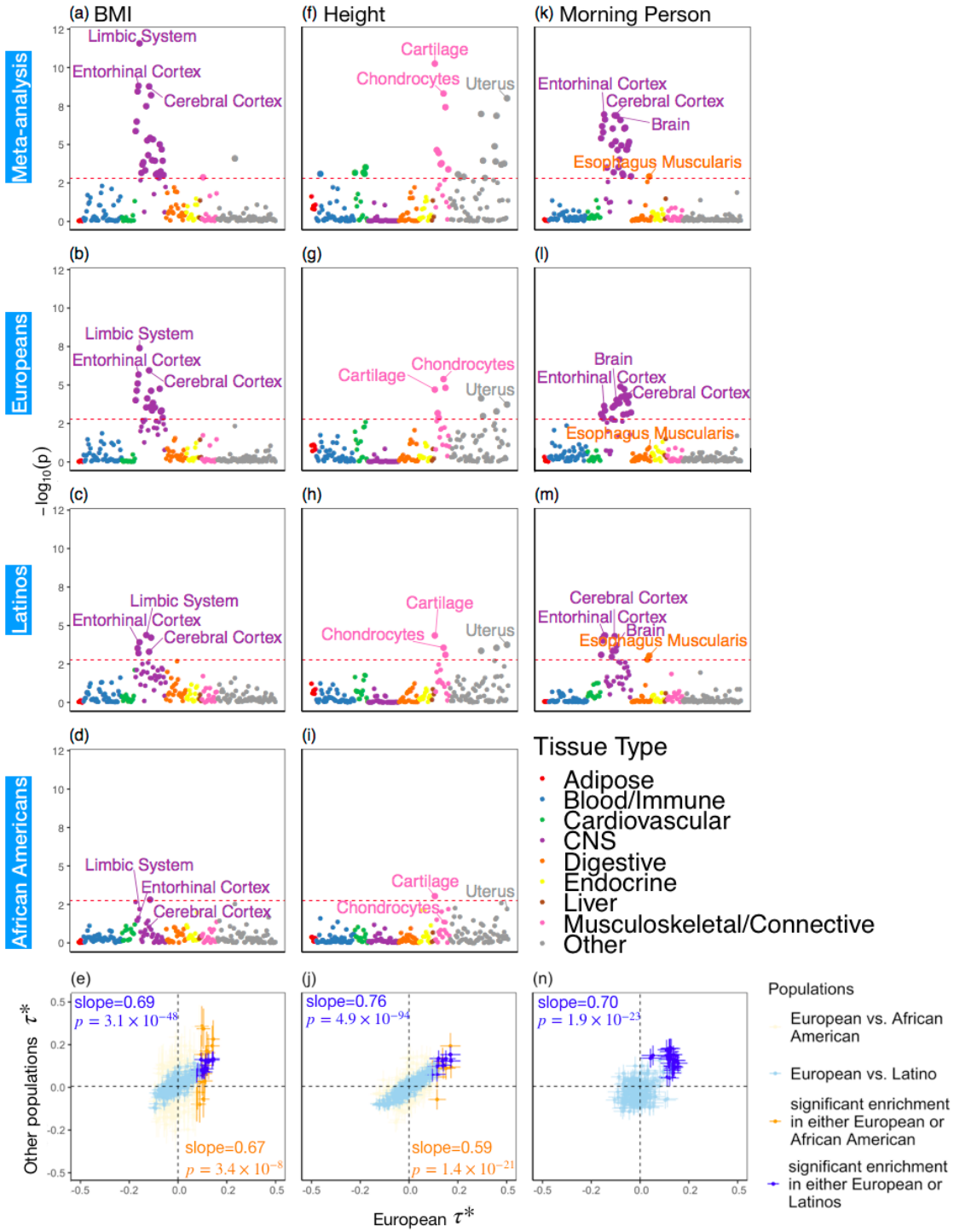
(a) BMI
(f) Height
(k) Morning Person
Meta-analysis
Europeans
Latinos
African Americans
Limbic System
Entorhinal Cortex
Cerebral Cortex
Cartilage
Chondrocytes
Uterus
Brain
Esophagus Muscularis
$-\log_{10}(p)$
Tissue Type
Adipose
Blood/Immune
Cardiovascular
CNS
Digestive
Endocrine
Liver
Musculoskeletal/Connective
Other
(e)
(j)
(n)
slope=0.69
$p = 3.1 \times 10^{-48}$
slope=0.67
$p = 3.4 \times 10^{-8}$
slope=0.76
$p = 4.9 \times 10^{-94}$
slope=0.59
$p = 1.4 \times 10^{-21}$
slope=0.70
$p = 1.9 \times 10^{-23}$
Other populations τ^*
European τ^*
Populations
European vs. African American
European vs. Latino
significant enrichment in either European or African American
significant enrichment in either European or Latinos

Table 1. h_g^2 **estimates of height, BMI and type 2 diabetes using different heritability estimation methods.** Reported values are estimates of h_g^2 (with standard deviations in brackets) from LDSC using a 20-cM window, cov-LDSC using a 20-cM window and 10 PCs, and GCTA using REAP to obtain the genetic relationship matrix with adjustment by 10 PCs. The final column provides reported h_g^2 estimates in European populations from various studies[10,32,33].

Phenotype	LDSC (baseline)	cov-LDSC (baseline)	GCTA (REAP w/ 10pc)	Public
Height	0.159 (0.037)	0.379 (0.079)	0.450 (0.042)	0.45-0.685[10,32]
BMI	0.113 (0.030)	0.248 (0.061)	0.235 (0.041)	0.246-0.27[32]
T2D	0.121 (0.035)	0.263 (0.073)	0.376 (0.046)	0.139-0.414[32,33]

801