Estimating heritability of complex traits in admixed populations with summary statistics

Yang Luo[1-5,*], Xinyi Li[1-5,*], Xin Wang[6], Steven Gazal[3,7], Josep Maria Mercader[3,8], 23andMe Research Team, 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 02115, 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 02115, USA
[11]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, 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).

All summary statistics-based methods to estimate the heritability of SNPs (h_g^2) rely on accurate linkage disequilibrium (LD) calculations. In admixed populations, such as African Americans and Latinos, LD estimates are influenced by admixture and can result in biased h_g^2 estimates. Here, we introduce covariate-adjusted LD score regression (cov-LDSC), a method to provide robust h_g^2 estimates from GWAS summary statistics and in-sample LD estimates in admixed populations. In simulations, we observed that unadjusted LDSC underestimates h_g^2 by 10%- 60%; in contrast, cov-LDSC is robust to all simulation parameters. We applied cov-LDSC to approximately 170,000 Latino, 47,000 African American 135,000 European individuals in three quantitative and five dichotomous phenotypes. Our results show that most traits have high concordance of h_g^2 between ethnic groups; for example in the 23andMe cohort, 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 of 0.10 ± 0.03, 0.33 ± 0.13 and 0.19 ± 0.01 in Latino, African American and European populations respectively.

Introduction

To date, genome-wide association studies (GWAS) have identified thousands of loci associated with hundreds of complex human traits and diseases[1]. However, the majority of GWAS, and the analytical tools developed to analyze GWAS data, have been focused on relatively homogenous continental populations, and in particular populations of European descent[2]. Non-European populations, particularly those with mixed ancestral background such as African Americans and Latinos, have been relatively understudied; diversifying GWAS data and analysis is important not only to ensure that the benefits of GWAS are shared beyond individuals of European ancestry but also because multi-population studies are valuable in detecting novel disease

associations, fine-mapping to causal variants, and exploring the extent to which the underlying genetic basis is shared across populations[3].

Investigators have developed statistical methods to estimate SNP-heritability (h_g^2), that is the proportion of phenotypic variance explained by genotyped variants, from GWAS data[4–6] . Summary statistics-based methods to estimate heritability, such as Linkage Disequilibrium score regression (LDSC)[5,7] and its extensions[5,7–9], have become particularly popular due to their computational efficiency, relative ease of application, and the requirement of only GWAS summary statistics rather than raw genotype data[10]. These methods have proven to be powerful tools in defining the genetic architecture of common traits[11], distinguishing polygenicity from confounding[5], establishing relationships between complex phenotypes[8] and defining key cell types and regulatory mechanisms of human diseases[7,12,13]. LDSC and other methods based on summary statistics, such as SumHer[14] implicitly rely on the assumptions that below a given distance threshold, typically set to one centimorgan (cM), in-sample LD can be well-approximated by reference panel LD, that beyond this threshold the in-sample LD between any two SNPs is independent of the distance between the SNPs and/or negligible, and that covariate adjustment does not have a large impact on in-sample LD. For studies of admixed populations, no reference panel has been shown to give a good representation of in-sample LD, LD continues to increase with distance well beyond 1-cM, and covariate adjustment has a large impact on in-sample LD. Thus, LDSC has not previously been applicable to admixed samples.

Results & Discussion

Overview of methods

In this work, we first examined the performance of LDSC in admixed populations and demonstrated that LDSC does indeed yield severely downward biased estimates of SNP-

heritability. Next, we extended the LDSC-based methods to admixed populations by introducing covariate-adjusted LDSC (cov-LDSC). Same as how summary statistics were computed, for each variant we regressed the global PCs, obtained within the GWAS samples, out of the raw genotype. LD scores were computed on the adjusted genotypes and used by LDSC to estimate heritability. Using covariate-adjusted in-sample LD to compute LD scores removes the issues of reference panel mismatch, long-distance admixture-LD, and covariate effects listed above, and produces robust estimates of heritability (**Method, Figure 1**).

We demonstrated that cov-LDSC is robust to a wide range of simulation scenarios. We then applied it to approximately 8,000 Latinos from the Slim Initiative in Genomic Medicine for the Americas (SIGMA) Type 2 Diabetes (T2D) Consortium[15] and approximately 162,000, 47,000 and 135,000 Latino, African Americans, and Europeans research participants, respectively, from the personal genetics company 23andMe. We analyzed three quantitative (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).

Robustness of LD score estimation

To demonstrate the effect of admixture on the stability of LD score estimates, we first calculated LD scores with genomic window sizes ranging from 0-50 cM in both European (EUR, N=503) and admixed American (AMR, N=347) populations from the 1000 Genomes Project[16]. As window size increases, we expect the mean LD score to reach a plateau because LD metrics should be negligible beyond a large enough genomic distance. If the mean LD score does not reach a plateau, but instead continues to increase with increasingly large window sizes, it may indicate one of two possibilities. Either (1) the window is too small to capture all of the LD or (2) the LD scores are capturing long-range pairwise SNP correlations arising from admixture; if this

increase is non-linear then there is non-negligible distance-dependent LD, violating LDSC assumptions. Examining unadjusted LD scores, we observed that in the EUR[5], the mean LD score estimates were stable with 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 1**). This suggests that cov-LDSC is able to correct the long-range LD due to admixture and yield stable estimates of LD scores (**Method, Supplementary Figure 1**), and also that cov-LDSC is applicable in homogeneous populations (**Supplementary Table 1**). The larger window size for the AMR population is 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, where the top two PCs capture 60.4% of variability in the data, LD score estimates are robust to different additional numbers of PCs and different window sizes 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[17]. 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 increases, $\hat{h}_g^2$ for LDSC increasingly underestimates 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%. Using cov-LDSC, while standard error increases with h_g^2, the point estimates remain 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 is also robust to confounding (**Supplementary Figure 3d**).

Finally, to assess the performance of cov-LDSC in polygenic binary phenotypes, we simulated studies of a binary trait with a prevalence of 0.1 using simulated genotypes and a liability threshold model (**Methods**). We showed that cov-LDSC provided robust 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.

Simulations with real genotypes

We next examined the performance of both unadjusted LDSC and cov-LDSC on real genotypes of individuals from admixed populations. We obtained data from the SIGMA cohort, which includes 8,214 Mexican and other Latino individuals. Using ADMIXTURE[18] and populations from the 1000 Genomes Project as reference panels, we observed that each individual in the SIGMA cohort has a varying degree of admixture proportion (**Supplementary Figure 5**). As in the AMR panel, we observed that using a 20-cM window, LD score estimates plateaued in SIGMA (**Supplementary Figure 6, Supplementary Table 2**), and were robust to different number of PCs (**Supplementary Figure 7**). We subsequently used a 20-cM window and 10 PCs in all simulations. We observed that cov-LDSC yielded unbiased estimates in traits with different polygenic genetic architectures by varying the number of causal variants and varying the total heritabilities (**Figure 2a-b**). In contrast LDSC underestimated heritability by as much as 62.5%. To examine the performance of cov-LDSC in the presence of environmental confounding factors, we simulated an environmental stratification component aligned with the first PC of the genotype data, representing European v.s. Native American ancestry. In this simulation scenario, cov-LDSC still provides unbiased h_g^2 estimates (**Figure 2c**). Intercepts of all the simulation scenarios are close to 1, suggesting that we had adequately controlled for confounding from population stratification and cryptic relatedness (**Supplementary Figure 8a-c**).

Thus far, we have used cov-LDSC by calculating LD scores on the same set of samples that were used for association studies (in-sample LD scores). In practical applications, computing LD scores on the whole data set can be computationally expensive and difficult to obtain, and so we investigated computing LD scores on a subset of samples. To investigate the minimum number of samples required to obtain accurate in-sample LD scores, we computed LD scores

178 on subsamples of 100, 500, 1,000 and 5,000 individuals from a GWAS of 10,000 simulated
179 genotypes. We also tested out-of-sample LD scores from 1,000 samples with a perfectly
180 matching demographic history in the simulated genotypes. cov-LDSC yielded unbiased
181 estimates for in-sample LD scores calculated using 1,000 samples (>10% of the total sample
182 size) and also using 1,000 samples in an out-of-sample reference panel with a perfectly
183 matching population structure (**Supplementary Figure 9**). We repeated these analyses in
184 simulated phenotypes in the SIGMA cohort. We subsampled the SIGMA chort, and obtained
185 unbiased estimates when using as few as 1,000 samples (**Figure 2d**). When using the AMR
186 panel as a reference panel for the SIGMA cohort, we observed an unbiased ${h_g}^2$ estimate (
187 $p = 0.33$, **Figure 2d**). This suggests that the AMR panel included in the 1000 Genomes Project
188 has similar demographic history compared to the SIGMA cohort (**Supplementary Figure 5**).
189 However, as the number of samples included in the subsampling decreased, the cov-LDSC
190 regression intercepts deviated further from 1 (**Supplementary Figure 8d**), probably due to
191 attenuation bias from noisily estimated LD scores at N<1,000. We therefore caution that when
192 using 1000 Genomes or any out-of-sample reference panels for a specific admixed cohort,
193 users should ensure that the demographic histories are shared between the reference and the
194 study cohort. We recommend computing in-sample LD scores on a randomly chosen subset of
195 at least 1,000 individuals from a GWAS.

196 ## Application to SIGMA and 23andMe cohorts

197 We next estimated ${h_g}^2$ of height, BMI and T2D phenotypes included in the SIGMA cohort of
198 8,214 samples and 943,244 variants (**Methods**) using cov-LDSC (**Table 1**). We estimated ${h_g}^2$
199 of height, BMI and T2D to be 0.38 ± 0.08 , 0.25 ± 0.06 and 0.26 ± 0.08 respectively. These
200 results are similar to what has been reported in the UK Biobank[19] and other studies[4,20] for
201 European populations. Although estimates differ in different studies (**Methods**), we noted that

without cov-LDSC, we would have obtained severely deflated estimates (**Table 1**). To confirm that our reported heritability estimates are robust under different model assumptions, we applied an alternative approach based on REML in the linear mixed model framework implemented in GCTA[21]. To avoid biases introduced from calculating genetic relatedness matrices (GRMs) in admixed individuals, we obtained a GRM based on an admixture-aware relatedness estimation method REAP[22] (Methods). GCTA-based results were similar to reported ${h_g}^2$ estimates from cov-LDSC, indicating our method is able to provide reliable ${h_g}^2$ estimates in admixed populations (**Table 1**).

We then applied both LDSC and cov-LDSC to 161,894 Latino, 46,844 African American and 134,999 European research participants from 23andMe, analyzing three quantitative and four dichotomous phenotypes (**Methods**). In this setting, using summary statistic methods to estimate heritability was essential since the dataset was too computationally expensive to apply genotype-based strategies. We used a 20-cM window and 10 PCs in LD score calculations for both populations (**Supplementary Figure 10**). LDSC and cov-LDSC produced similar heritability estimates in the European population, whereas in the admixed populations, LDSC consistently provided low estimates of ${h_g}^2$ (**Supplementary Table 3**). For each phenotype, we estimated ${h_g}^2$ using the same population-specific in-sample LD scores. For most phenotypes, the reported ${h_g}^2$ is similar among the three ethnic groups with a notable exception for age at menarche (**Figure 3**), suggesting possible differences ($p = 7.1\times10^{-3}$ 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[23,24]. Early menarche might influence the genetic architecture 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[25,26]. These results highlight the importance of

227 including diverse populations in genetic studies in order to enhance our understanding of
228 complex traits that show differences in their genetic heritability.

229 Conclusion

230 As we expand genetic studies to explore admixed populations around the world, extending
231 statistical genetics methods to make inferences within admixed populations is crucial. This is
232 particularly true for methods based on summary statistics, which are dependent on the use of
233 LD scores, which we showed to be problematic in admixed populations. In this study, we
234 demonstrated that original LDSC and other summary statistics-based methods, such as
235 PCGCs[27] and SumHer[28], that were originally designed for homogenous populations, potentially
236 severely underestimated heritability in admixed populations. We introduced cov-LDSC which
237 regresses out global PC on individual genotypes during the LD score calculation, and showed it
238 can yield robust LD score and heritability estimates in both homogenous and admixed
239 populations.
240
241 By applying cov-LDSC to Europeans, African Americans, and Latin Americans in the 23andMe
242 cohort, we observed evidence of heritability differences across different populations. These
243 differences highlight the importance of studying diverse populations. How these differences may
244 correspond to differences in biological mechanisms may lead to mechanistic insights about the
245 phenotype. One strategy to do this, which we will explore in the future is to extend cov-LDSC to
246 partition heritability by different functional annotations and cell types to dissect the genetic
247 architecture in admixed populations.
248
249 Although our work provided a novel approach to estimate genetic heritability using summary
250 statistics in admixed populations, it has a few limitations (**Methods**). First, covariates included in
251 the summary statistics should match the covariates included in the covariate-adjusted LD score

calculations (**Supplementary Figure 11**), and h_g^2 estimates in admixed populations are more sensitive to their matching LD reference panels. Unmatched reference panels are likely to produce biased estimates[29,30]. 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 GWAS summary statistics when possible. 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[31,32]. 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. And any extension to a larger set of genetic variants, especially across different cohorts should be performed with caution. Third, recent studies have shown that heritability estimates can be sensitive to the choice of the frequency-dependent heritability model[6,9,14]. However, this is unlikely to impact the main conclusions of the current study[9] and how to incorporate ancestry-dependent frequencies in the LD-dependent annotation remains a subject of future study (**Methods**).

Despite these limitations, in comparison with other methods, such as those based on restricted maximum likelihood estimation (REML)[21] 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 only need to be calculated once per cohort and can be obtained with a subset of samples. This is particularly useful in large cohorts such as 23andMe and UK Biobank[33], where multiple phenotypes have been collected per individual. In this setting, 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 may be extended to perform analyses such as estimating genetic

277 correlations, partitioning $h_g{}^2$ by functional annotations, identifying disease-relevant tissues and

278 cell types and multi-trait analysis[7,34,35].

279

280 Methods

281 Mathematical framework

282 The main LD score regression[5] draws a linear model between χ^2 statistics and heritability $h_g{}^2$:

$$E[\chi^2{}_j] = \frac{N h_g{}^2}{M} l_j + Na + 1$$

283

284 where N is number of samples; M is number of SNPs; a measures the confounding biases

285 including cryptic relatedness and population stratification; and l_j is LD score of variant j.

286 measured as:

287 $$\ell_j = \sum_{k=1}^{M} r_{jk}^2 .$$

288 Let X_{ij} be the genotype of individual i at SNP j, standardized that $\widehat{X_{ij}}$ has mean 0 variance 1 for

289 each SNP. In the original LD score regression, the in-sample correlation $\hat{r}_{jk}$ between SNP j and

290 SNP k is defined as:

291 $$\hat{r}_{jk} = \frac{1}{N} \hat{X}_j^T \hat{X}_k .$$

292 We introduced cov-LDSC for admixed populations. The intuition of cov-LDSC is to regress out

293 the ancestral or any other fix effects for each SNP j from its genotype. Define C_i as the covariate

294 or top principal components of individual i. We adjusted the standardized $\hat{X}$matrix to X' by

295 continuously subtracting the projection of covariate from raw genotypes

296 $X' = \hat{X} - CC^T \hat{X}$.

297 We then standardized X' to be mean 0 and variance 1 for each SNP, denoted as $\hat{X}'$. Based on
298 the adjusted genotypes $\hat{X}'$, we measure in-sample cov-LD score $\hat{r}'_{jk}$ in admixed populations:

299 $$\hat{r}'_{jk} = \frac{1}{N}\hat{X}'^{T}_{j}\hat{X}'_{k}$$

300 Window size and number of PCs in LD score calculations

301 To determine the optimal window size for estimating LD scores, we examined the effect of
302 varying the genomic widow size for both simulated and real data sets. We concluded that LD
303 score estimates were robust to the choice of window size if the increase in the mean LD score
304 estimates was less than 1% per cM beyond a given window. Using this criterion, we used
305 window sizes of 5-cM and 20-cM for the simulated and real genotypes respectively
306 (**Supplementary Table 2, 4-5**). We also calculated the squared correlations between LD score
307 estimates using the chosen window size and other LD score estimates with window sizes larger
308 than the chosen window. The squared correlations were greater than 0.99 in all cases
309 (**Supplementary Table 6-8**) indicating the LD score estimates were robust at the chosen
310 window sizes.
311
312 Similarly, to determine the number of PCs needed to be included in the GWAS association tests
313 and cov-LDSC calculations, we examined the effect of varying the genomic window size using
314 different numbers of PCs. The number of PCs that needs to be included for covariate
315 adjustment depends on the population structure for different datasets. In practice, we
316 recommend using the same number of PCs to adjust for the GWAS association tests and for LD
317 score calculations (**Supplementary Figure 11)**.
318

319 Genotype simulations

We used msprime[17] version 0.6.1 to simulate population structure with mutation rate 2×10^{-8} and recombination maps from the HapMap Project[36]. The demographic model was adapted from Mexican migration history[37] and the parameters were previously inferred from the 1000 Genomes Project[16] . We assumed the admixture event happened approximately 500 years ago to mirror the European colonization of the Americas. We set different admixture proportions to reflect different admixed populations. In each population, 10,000 individuals were simulated after removing second degree related samples (kinship>0.125) using KING[38].

We applied 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.

Phenotype simulations

We used two phenotype simulation strategies implemented in the GCTA[21] and the baseline model[7] respectively. These two strategies assume different genetic architectures of complex traits. In the GCTA model, all variants are equally likely to be causal independent of their functional or minor allele frequency (MAF) structure. On the other hand the baseline model incorporates functionally dependent architectures. Briefly, it includes 53 annotations overlapping genome-wide functional annotations (e.g. coding, conserved, regulatory). All causal variants were generated among common observed variants with MAF >5% (~40,000 SNPs in simulated genotypes and 943,244 SNPs in SIGMA cohort).

Both models assume an additive genetic model $Y_j = W_{ij}\beta_i + \epsilon_j$, where Y_j is the phenotype for the j^{th}individual; $W_{ij} = \frac{X_{ij}-2p_i}{\sqrt{2p_i(1-2p_i)}}$ is the standardized genotype of X_{ij} for the i^{th} causal variant (with MAF ≥ 5%) of thej^{th} individual and p_i being the frequency of the i^{th} causal variant. β_j is the allelic effect of the standardized genotype of the j^{th} causal variant and ϵ_j is the residual

effect generated from a normal distribution with mean 0 and variance $\sigma_e{}^2$. In the GCTA model, the standardized casual effect size variance is constant, i.e. $var(\beta_j) = h_g{}^2/M$, whereas in the baseline model $var(\beta_j) = \sum_C a_c\,(j)\,\tau_c$, where $a_c\,(j)$ is the value of annotation a_c at variant j and τ_c represents the per-variant contribution, of one unit of the annotation a_c , to heritability.

We used recommended parameters in both strategies and applied it in all simulation scenarios in the SIGMA cohort and observed no significant differences in heritability estimates (**Supplementary Table 9**). In all simulations, we restricted ourselves to genotyped SNPs with MAF ≥ 5% as recommended in previous studies[6,7]. We concluded that the total genetic heritability estimated using cov-LDSC is robust under both models in all simulation scenarios.

For case-control simulations, we adopted a liability threshold model with disease prevalence 0.1. 5,000 cases and 5,000 controls were obtained for each simulation scenario. To represent environmental stratification, similar to previously described[5], we added 0.2 * standardized first principal component to the standardized phenotypes. .

LD score estimates

We calculated in-sample LD scores using both a non-stratified LD score model and baseline model[7]. We used the 53 non-frequency dependent annotations included the baseline model to estimate $h_g{}^2$ in the 23andMe research database and the SIGMA cohort. $h_g{}^2$ estimates of three quantitative traits and five binary traits were robust when using different LD models (**Supplementary Table 3**). We recognized recent studies have shown that genetic heritability can be sensitive to the choice of LD-dependent heritability model[6,9]. However, in the admixed population, it is complicated to create LD-related annotations that are independent from admixture-LD. We would need a larger and denser admixed sequencing panel to evaluate the

368 performance of baseline-LD model in admixed populations. Regardless, this should not impact
369 the result of current study, where we reported the total phenotypic variation that can be
370 explained by common HapMap3 variants[9].
371

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

374 8,214 Mexican and other Latin American samples were genotyped with Illumina HumanOmni2.5
375 array. The genotyped data were pre-phased using SHAPEIT2[39]. IMPUTE2[40] was then used to
376 impute genotypes at untyped genetic variants using the 1000 Genomes Project Phase 3[16]
377 dataset as a reference panel. We merged genotyped SNPs and imputed variants with INFO
378 >0.99. Merged sets of SNPs were further filtered to be HapMap3 variants with MAF >5% and
379 SNPs in high LD regions were removed. After QC, 8,214 individuals and 943,244 SNPs
380 remained. We examined three phenotypes from the SIGMA cohort: height, BMI, and type 2
381 diabetes. For each phenotype, we included age, sex, and the first 10 PCs as fixed effects in the
382 association analyses.
383
384 We removed high LD regions (**Supplementary Table 10**) and used a 20-cM window and 10
385 PCs in all scenarios. h_g^2 estimates were robust at a 20-cM window with 10 PCs when using
386 cov-LDSC (**Supplementary Figure 12**) with an assumed prevalence of 0.144[15]. Intercepts of all
387 described simulated scenarios are shown in **Supplementary Figure 8.**

388 23andMe cohort

389 All participants were drawn from the customer base of 23andMe, Inc., a direct to consumer
390 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 were then chosen from consented individuals who were genotyped successfully on the v5 platform, an Illumina Infinium Global Screening Array (~640,000 SNPs) supplemented with ~50,000 SNPs of custom content. Participants were restricted to a set of individuals who have European, African American, or Latino ancestry determined through an analysis of local ancestry[41].

To compute LD scores, both genotyped and imputed SNPs were used. Genotype variants were filtered to have genotype call rate > 90%, self-chain score = 0, further restricted to eliminate those with strong evidence of Hardy Weinberg disequilibrium ($p > 10^{-20}$), and passing a parent-offspring transmission test. Imputed variants used a reference panel that combined the May 2015 release of the 1000 Genomes Phase 3 haplotypes[16] with the UK10K imputation reference panel[42]. Imputed dosages were rounded to the nearest integer (0, 1, 2) for downstream analysis. Variants were filtered to have imputation r-squared > 0.9. Both genotyped and imputed variants were also filtered for batch effects and sex dependent effects. To minimize rounding inaccuracies, genotyped SNPs were prioritized over imputed SNPs in the merged SNP set. The merged SNP set were further restricted to HapMap3 variants with MAF $\geq$ 0.05. We have measured 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 5**) and squared correlation between different window sizes (**Supplementary Table 8)**. All LD scores were computed with a 20-cM window.

In genome-wide association analyses, for each population, a maximal set of unrelated individuals was chosen for each analysis using a segmental identity-by-descent (IBD) estimation algorithm[43]. Individuals were defined as related if they shared more than 700-cM IBD.

All association tests were performed using linear regression model for quantitative traits and logistic regression model for binary traits assuming additive allelic effects. We included covariates for age, sex and the top 10 PCs to account for residual population structure. Details of phenotypes and genotypes are listed in **Supplementary Table 11.**

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 ≥5%, including tagged causal effects of both low-frequency and common variants. This quantity is different from h_{common}^2, the heritability causality explained byall common SNPs excluding tagged causal effects of low-frequency variants, reported in the original LDSC[5]. When applying LDSC to Europeans and other homogeneous populations, it is recommended to use an appropriate sequenced reference panel, such as 1000 Genome Project, which includes >99% of the SNPs with frequency >1%[16], which allows for the estimation of h_{common}^2. 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 recommend to use common HapMap3 SNPs (MAF ≥ 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 used all well imputed (INFO>0.99, **Methods**) SNPs (~6.9 million) in SIGMA cohort as 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 12**). The difference that we observed is consistent with the previous discovery that the low frequency variants (0.5%<MAF<5%) explains $6.3 \pm 0.2\times$ (15.87%) of heritability less than common variants (MAF≥5%) on average[44].

URLs

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. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* **101,** 5–22 (2017).
2. Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. *Nature* **538,** 161–164 (2016).
3. Seldin, M. F., Pasaniuc, B. & Price, A. L. New approaches to disease mapping in admixed populations. *Nat. Rev. Genet.* **12,** 523–528 (2011).
4. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42,** 565–569 (2010).
5. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47,** 291–295 (2015).
6. Speed, D. *et al.* Reevaluation of SNP heritability in complex human traits. *Nat. Genet.* **49,** 986–992 (2017).
7. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47,** 1228–1235 (2015).
8. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47,** 1236–1241 (2015).
9. Gazal, S. *et al.* Linkage disequilibrium-dependent architecture of human complex traits

shows action of negative selection. *Nat. Genet.* **49,** 1421–1427 (2017).

10. Pasaniuc, B. & Price, A. L. Dissecting the genetics of complex traits using summary association statistics. *Nat. Rev. Genet.* **18,** 117–127 (2017).
11. Evans, L. M. *et al.* Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits. *Nat. Genet.* **50,** 737–745 (2018).
12. Finucane, H. K. *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* **50,** 621–629 (2018).
13. Guo, J., Yang, J. & Visscher, P. M. Leveraging GWAS for complex traits to detect signatures of natural selection in humans. *Curr. Opin. Genet. Dev.* **53,** 9–14 (2018).
14. Speed, D. & Balding, D. J. SumHer better estimates the SNP heritability of complex traits from summary statistics. *Nat. Genet.* (2018). doi:10.1038/s41588-018-0279-5
15. SIGMA Type 2 Diabetes Consortium *et al.* Sequence variants in SLC16A11 are a common risk factor for type 2 diabetes in Mexico. *Nature* **506,** 97–101 (2014).
16. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).
17. Kelleher, J., Etheridge, A. M. & McVean, G. Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLoS Comput. Biol.* **12,** e1004842 (2016).
18. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19,** 1655–1664 (2009).
19. Ge, T., Chen, C.-Y., Neale, B. M., Sabuncu, M. R. & Smoller, J. W. Phenome-wide heritability analysis of the UK Biobank. *PLoS Genet.* **13,** e1006711 (2017).
20. Xue, A. *et al.* Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. *Nat. Commun.* **9,** 2941 (2018).
21. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88,** 76–82 (2011).
22. Thornton, T. *et al.* Estimating Kinship in Admixed Populations. *Am. J. Hum. Genet.* **91,**

122–138 (2012).

23. Demerath, E. W. *et al.* Genome-wide association study of age at menarche in African-American women. *Hum. Mol. Genet.* **22,** 3329–3346 (2013).
24. Fernández-Rhodes, L. *et al.* 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 (2018).
25. Canoy, D. *et al.* Age at menarche and risks of coronary heart and other vascular diseases in a large UK cohort. *Circulation* **131,** 237–244 (2015).
26. Bodicoat, D. H. *et al.* Timing of pubertal stages and breast cancer risk: the Breakthrough Generations Study. *Breast Cancer Res.* **16,** R18 (2014).
27. Weissbrod, O., Flint, J. & Rosset, S. Estimating SNP-Based Heritability and Genetic Correlation in Case-Control Studies Directly and with Summary Statistics. *Am. J. Hum. Genet.* **103,** 89–99 (2018).
28. Speed, D. & Balding, D. Better estimation of SNP heritability from summary statistics provides a new understanding of the genetic architecture of complex traits. *bioRxiv* 284976 (2018). doi:10.1101/284976
29. Yang, J. *et al.* Genome-wide genetic homogeneity between sexes and populations for human height and body mass index. *Hum. Mol. Genet.* **24,** 7445–7449 (2015).
30. Yang, J., Zeng, J., Goddard, M. E., Wray, N. R. & Visscher, P. M. Concepts, estimation and interpretation of SNP-based heritability. *Nat. Genet.* **49,** 1304–1310 (2017).
31. Martin, A. R. *et al.* Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *Am. J. Hum. Genet.* **100,** 635–649 (2017).
32. Nelson, S. C. *et al.* 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 (2016).

541 33. Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide
542 range of complex diseases of middle and old age. *PLoS Med.* **12,** e1001779 (2015).
543 34. Turley, P. *et al.* Multi-trait analysis of genome-wide association summary statistics using
544 MTAG. *Nat. Genet.* (2018). doi:10.1038/s41588-017-0009-4
545 35. Gusev, A. *et al.* Partitioning heritability of regulatory and cell-type-specific variants across
546 11 common diseases. *Am. J. Hum. Genet.* **95,** 535–552 (2014).
547 36. International HapMap Consortium. The International HapMap Project. *Nature* **426,** 789–796
548 (2003).
549 37. Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H. & Bustamante, C. D. Inferring the
550 joint demographic history of multiple populations from multidimensional SNP frequency
551 data. *PLoS Genet.* **5,** e1000695 (2009).
552 38. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies.
553 *Bioinformatics* **26,** 2867–2873 (2010).
554 39. O'Connell, J. *et al.* A general approach for haplotype phasing across the full spectrum of
555 relatedness. *PLoS Genet.* **10,** e1004234 (2014).
556 40. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and
557 accurate genotype imputation in genome-wide association studies through pre-phasing.
558 *Nat. Genet.* **44,** 955–959 (2012).
559 41. Durand, E. Y., Do, C. B., Mountain, J. L. & Michael Macpherson, J. Ancestry Composition:
560 A Novel, Efficient Pipeline for Ancestry Deconvolution. *bioRxiv* 010512 (2014).
561 doi:10.1101/010512
562 42. UK10K Consortium *et al.* The UK10K project identifies rare variants in health and disease.
563 *Nature* **526,** 82–90 (2015).
564 43. Henn, B. M. *et al.* Cryptic distant relatives are common in both isolated and cosmopolitan
565 genetic samples. *PLoS One* **7,** e34267 (2012).
566 44. Gazal, S. *et al.* Functional architecture of low-frequency variants highlights strength of

negative selection across coding and non-coding annotations. *Nat. Genet.* **50,** 1600–1607 (2018).

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 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.

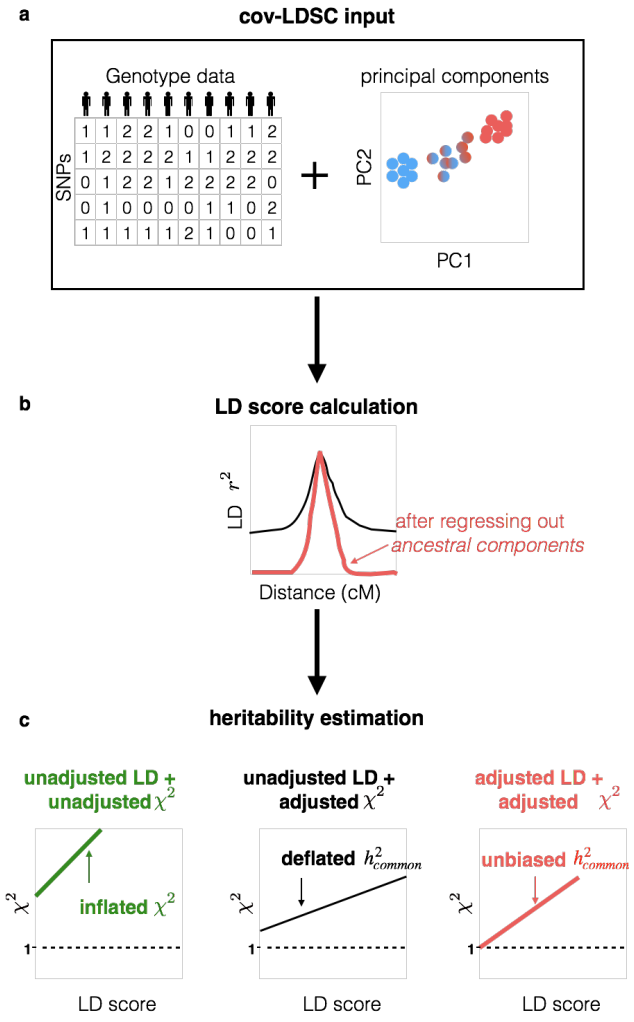

Figure 2. Estimates of heritability (h_g^2) under different simulation scenarios using the SIGMA cohort. LDSC (orange) underestimated h_g^2 and cov-LDSC (blue) yielded robust h_g^2 estimates under all settings. Each boxplot represents the mean LD score estimate from 100 simulated phenotypes using the 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 ann 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.

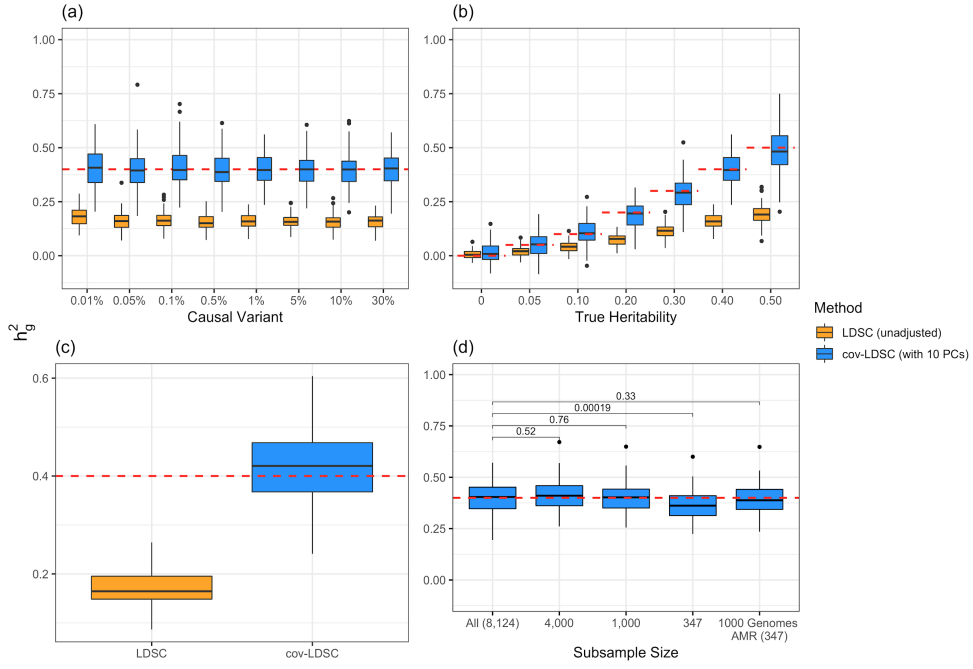

Figure 3. Estimates of heritability (h_g^2) of three quantitative and four dichotomous traits in two admixed population in the 23andMe research cohort. For seven selected non-disease phenotypes (body mass index (BMI), height, age at menarche, left handedness, morning person, motion sickness and nearsightedness) 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 sample size used in obtained summary statistics used in cov-LDSC. For BMI and height, we also reported the h_g^2 estimates from the SIGMA cohort.

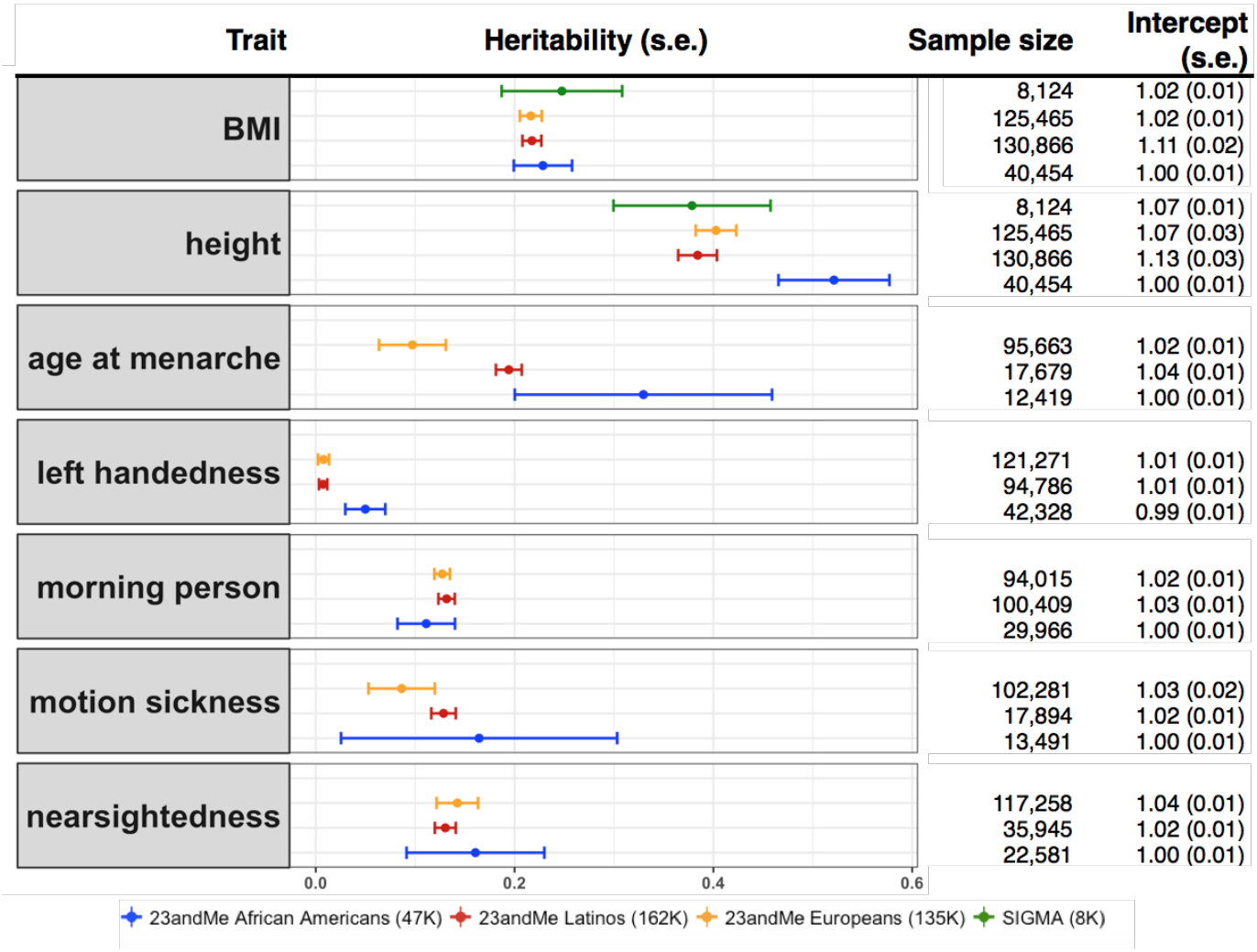

Table 1. h_g^2 estimates of height, body mass index (BMI) and type 2 diabetes (T2D) using different heritability estimation methods. Reported values are estimates of h_g^2 (with standard errors in brackets) from LDSC using a 20-cM window, cov-LDSC using a 20-cM window and 10 PCs, and GCTA using REAP 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[4,19,20].

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[4,19]
BMI	0.113 (0.030)	0.248 (0.061)	0.235 (0.041)	0.246-0.27[19]
T2D	0.121 (0.035)	0.263 (0.073)	0.376 (0.046)	0.139-0.414[19,20]