Published in final edited form as:
Nat Genet. 2019 February ; 51(2): 277–284. doi:10.1038/s41588-018-0279-5.

SumHer better estimates the SNP heritability of complex traits from summary statistics

Doug Speed[1,2,3,*] **David J Balding**[3,4]

[1]Aarhus Institute of Advanced Studies (AIAS), Aarhus University, Denmark

[2]Bioinformatics Research Centre, Aarhus University, Denmark

[3]UCL Genetics Institute, University College London, United Kingdom

[4]Melbourne Integrative Genomics, School of BioSciences and School of Mathematics & Statistics, University of Melbourne, Australia

Abstract

We present SumHer, software for estimating confounding bias, SNP heritability, enrichments of heritability and genetic correlations using summary statistics from genome-wide association studies. The key difference between SumHer and the existing software LD Score Regression (LDSC) is that SumHer allows the user to specify the heritability model. We apply SumHer to results from 24 large-scale association studies (average sample size 121,000) using our

*Corresponding author: doug@aias.au.dk.

Data availability

The simulations and 25 raw GWAS used data from the Wellcome Trust and eMERGE Network. These were applied for and downloaded from, respectively, the European Genome-phenome Archive (accession codes: EGAD00000000001, EGAD00000000002, EGAD00000000003, EGAD00000000004, EGAD00000000005, EGAD00000000006, EGAD00000000007, EGAD00000000008, EGAD00000000009, EGAD00000000021, EGAD00000000022, EGAD00000000023, EGAD00000000024, EGAD00000000025, EGAD00000000057, EGAD00010000124, EGAD00010000264, EGAD00010000506, EGAD00010000634, EGAS00001000672) and from dbGaP (accession codes: phs000888.v1.p1.c1, phs000888.v1.p1.c3, phs000888.v1.p1.c4, phs000888.v1.p1.c5). To investigate the impact of the reference panel, we used data from the Health and Retirement Study, also available from dbGaP (accession code: phs000428.v2.p2). Results for each of the 24 summary GWAS are available to download from the websites of the corresponding studies (see Table 1 for references).

Accession Codes

Wellcome Trust Case Control Consortium data: EGAD00000000001, EGAD00000000002, EGAD00000000003, EGAD00000000004, EGAD00000000005, EGAD00000000006, EGAD00000000007, EGAD00000000008, EGAD00000000009, EGAD00000000021, EGAD00000000022, EGAD00000000023, EGAD00000000024, EGAD00000000025, EGAD00000000057, EGAD00010000124, EGAD00010000264, EGAD00010000506, EGAD00010000634, EGAS00001000672. eMERGE Network data: phs000888.v1.p1.c1, phs000888.v1.p1.c3, phs000888.v1.p1.c4, phs000888.v1.p1.c5. Health and Retirement Study data: phs000428.v2.p2.

Code availability

Step-by-step code for using SumHer to estimate SNP heritability, confounding bias, heritability enrichments and genetic correlations from summary statistics are provided in the Supplementary Note.

URLs

LDAK, http://www.ldak.org; LDSC, http://www.github.com/bulik/ldsc; 24 functional annotations, https://data.broadinstitute.org/alkesgroup/LDSCORE/baseline_bedfiles.tgz.

Author contributions

D.S. performed the analysis, D.S. and D.J.B. wrote the manuscript.

Competing interests

The authors declare no competing financial interests.

Europe PMC Funders Author Manuscripts

recommended heritability model. We show that these studies tended to substantially over-correct for confounding, and as a result the number of genome-wide significant loci was under-reported by about a quarter. We also estimate enrichments for 24 categories of SNPs defined by functional annotations. A previous study using LDSC reported that conserved regions were 13-fold enriched, and found a further six categories with above 3-fold enrichment. By contrast, our analysis using SumHer finds that none of the categories have enrichment above 2-fold. SumHer provides an improved understanding of the genetic architecture of complex traits, which enables more efficient analysis of future genetic data.

LD Score Regression (LDSC) has been frequently applied to summary statistics from genome-wide association studies (GWAS).1–4 It has four main aims: to estimate the average bias due to confounding, to estimate the "SNP heritability" of a trait (the proportion of phenotypic variance explained by all SNPs), to estimate the heritability enrichments of SNP categories, and to estimate the genetic correlation between a pair of traits. LDSC assumes a specific heritability model in which each SNP in the genome is expected to contribute equally.1 Although this model is widely used in statistical genetics, we recently showed that across a range of human traits, it poorly reflects reality.5 In particular, it fails to appreciate that in regions of high linkage disequilibrium (LD), the average heritability of each SNP tends to be lower due to multiple tagging of causal variation.6 As a result of this model misspecification, LDSC tends to over-estimate confounding bias, under-estimate SNP heritability and produce exaggerated estimates of enrichments.

We propose SumHer, software for estimating SNP heritability from summary statistics that allows the user to specify the heritability model. We apply SumHer to publicly-available GWAS results for 24 disease and quantitative traits, using our recommended heritability model.5,6 Firstly we show that these GWAS tended to over-correct for confounding; although in total they reported 2060 genome-wide significant loci, were it not for this over-correction they would have discovered approximately 2800. Secondly, we show that for each of the 24 traits, the SumHer estimate of SNP heritability is at least double that from LDSC. Thirdly, we consider functional enrichments. A previous study using LDSC concluded that heritability is highly concentrated in specific functional categories;2 for example, it estimated that across 17 diseases, conserved regions contribute 35% of SNP heritability, indicating that they are 13-fold enriched for causal variation. When we repeat the analysis using SumHer and our 24 traits, the concentration of heritability in functional categories is more modest: for example, while conserved regions remain significantly enriched, we estimate that they contribute only 7.1% (s.d. 0.2) of SNP heritability, representing 2.0-fold (s.d. 0.07) enrichment. We finish by providing an example of how results from SumHer enable more efficient analysis of genetic data. We show how for body mass index, height, HDL & LDL cholesterol and triglyceride, we are able to significantly improve the predictive performance of polygenic risk scores by incorporating our heritability model and estimates of enrichments. We make SumHer freely available within our software package LDAK (www.ldak.org).6

Europe PMC Funders Author Manuscripts

Results

SumHer

SumHer has the same four aims as LDSC;1–3 we outline them here, with methodological details provided in Online Methods. Suppose we are provided with summary statistics from a GWAS where each of m SNPs has been tested individually for association with a particular trait. Suppose also that we have access to a well-matched reference panel, from which we can reliably estimate r^2_{jl} the squared correlation between SNPs j and l. Let S_1, S_2,…,S_m denote the $\chi^2(1)$ test statistics from single-SNP analysis; the first aim is to estimate the average inflation of these test statistics due to confounding. Let h^2_j denote the heritability directly contributed by SNP j; the second aim is to estimate $h^2_{SNP} = \sum_j h^2_j$ the SNP heritability of the trait.7 Let $\mathbb{C}$ index a category of SNPs; the third aim is to estimate $(\sum_{j\in\mathbb{C}} h^2_j) / h^2_{SNP}$, the share8 of h^2_{SNP} contributed by SNPs in $\mathbb{C}$ (we can then estimate the enrichment of the category by dividing its estimated share by its expected share). Finally, if we are also provided with summary statistics from a second trait, the fourth aim is to estimate the correlation between SNP effect sizes for the two traits.9

In order to achieve these four aims, we require a heritability model, which describes how h^2_j is expected to vary across the genome. Suppose this heritability model takes the form $E[h^2_j] \propto q_j$. The main difference between LDSC and SumHer is that SumHer allows for any heritability model (i.e., the user can specify arbitrary q_j), whereas LDSC assumes all q_j are the same. We recommend using SumHer with the "LDAK Model":

$$q_j = \left[f_i\left(1 - f_j\right)\right]^{0.75} w_j, \quad (1)$$

where f_j is the minor allele frequency (MAF) of SNP j and w_j is a weighting based on local levels of LD.5,6 In this model, a SNP with high MAF is expected to contribute more heritability than one with low MAF, while a SNP in a region of low LD is expected to contribute more than one in a region of high LD. By contrast, LDSC estimates are obtained by setting $q_j = 1$, which corresponds to the assumption that all SNPs are expected to contribute equally.1 We refer to this as the "GCTA Model" as it is a core assumption of the software GCTA.5,10

A second difference between LDSC and SumHer is how they estimate confounding bias. In a GWAS with no confounding, $E[S_j] \approx 1 + nv^2_j$, where n is the sample size and $v^2_j = h^2_j + \sum_{l \text{ near } j} r^2_{jl} h^2_l$ is the heritability tagged by SNP j (a working definition of "near" is within 1 Mb[1]). Both LDSC and SumHer estimate the deviation of test statistics from their expected values assuming no confounding. LDSC uses the model $E[S_j] \approx 1 + A + nv^2_j$, where A indicates the average amount test statistics are inflated additively due to confounding (LDSC then reports 1 + A, which it refers to as "the intercept").1 By contrast, we recommend using the model $E[S_j] \approx C(1 + nv^2_j)$, where C is the average amount test statistics are inflated multiplicatively. While switching between additive and multiplicative models of inflation will not generally impact estimates of confounding bias, enrichment or genetic correlation, it will affect estimates of h^2_{SNP} (see Online Methods). We prefer to model inflation as

multiplicative because our analyses below indicate that the largest contributor of bias in published GWAS is genomic control, which affects test statistics multiplicatively.11

In total we use six versions of SumHer, which differ according to their assumed heritability and confounding models. LDSC-Zero assumes the GCTA Model and that test statistics are not inflated due to confounding (A = 0, C = 1); this is equivalent to using the LDSC software[1] with the option "--intercept-h2 1". LDSC assumes the GCTA Model and that confounding inflation is additive (A free to vary, C = 1); this is equivalent to using the LDSC software[1] with default options. SumHer-Zero assumes the LDAK Model and that there is no confounding inflation (A = 0, C = 1); this is our recommended version when estimating h^2_{SNP} or enrichments and confident that confounding is negligible. SumHer-GC assumes the LDAK Model and that confounding inflation is multiplicative (A = 0, C free to vary); this is our recommended version when estimating confounding bias or genetic correlation, or when estimating h^2_{SNP} or enrichments and it is likely that test statistics are inflated due to population structure or familial relatedness, or were obtained using genomic control or mixed-model association analysis (see below). Hybrid-Zero assumes the heritability model

$$q_j = (1-p) \times 1/m + p \times \left[f_i\left(1-f_j\right)\right]^{0.75} w_j / Q' \quad \text{where} \quad Q' = \sum_j \left[f_j(1-f_j)\right]^{0.75} w_j, \qquad (2)$$

and that there is no confounding inflation (A = 0, C = 1), while **Hybrid-GC** assumes the same heritability model and that confounding inflation is multiplicative (A = 0, C free to vary). Model (2) is a linear combination of the GCTA and LDAK models, where p indicates the weight assigned to the LDAK model; we use this heritability model to compare the fit of the GCTA and LDAK models on real data (see below).

For all analyses, we consider only common (MAF ≥ 0.01), biallelic, autosomal SNPs. Following previous guidelines,1,2,4,5 we exclude SNPs within the major histocompatibility complex (MHC; Chromosome 6: 25-34 Mb), as well as SNPs which individually explain > 1% of phenotypic variation, and SNPs in LD with these (within 1 cM and $r^2_{jl} > 0.1$). Our reference panel is 404 non-Finnish, European individuals from the 1000 Genomes Project, recorded for 8,598,995 SNPs.12 For our enrichment analyses, we consider the same 24 functional annotations as Finucane *et al.*[2] which include coding, conserved, enhancer and promoter regions (see Supplementary Table 1 for a full list). When estimating enrichments using LDSC-Zero or LDSC, we use the 53-part model recommended by Finucane *et al.*2 (one category for each annotation, one for each of 28 "buffer regions", plus one containing all SNPs); when using SumHer-Zero or SumHer-GC, we use a 25-part model (one category for each annotation, plus one containing all SNPs).

Summary statistics

For our real analyses, we use summary statistics from two sets of GWAS. The "25 raw GWAS" (18 binary traits, 7 quantitative, average sample size 9 700; see Supplementary Table 2) are those for which we have access to individual-level data, from either the Wellcome Trust Case Control Consortium13 (WTCCC) or the eMERGE Network,14 and therefore we perform the association analysis ourselves. The "24 summary GWAS" (9

binary traits, 15 quantitative, average sample size 121,000; see Table 1), are those for which we use summary statistics from previously-published analyses.15

For each of the 25 raw GWAS, we perform strict quality control (see Online Methods), then use REML to estimate how much of the total phenotypic variance explained by SNPs is inflation due to population structure or familial relatedness.8,16 We estimate that on average 3.1% (range 0.2% to 7.2%) of the variance explained is inflation (Supplementary Fig. 1), indicating that confounding due to population structure or familial relatedness is is modest. For the 24 summary GWAS, we are largely reliant on the quality control decisions of the original authors, which are generally much less strict than ours (Supplementary Table 3). For 4 of the GWAS, imputation info scores are available, so we exclude SNPs with score < 0.95; for the remaining 20 GWAS, we instead restrict to the 4,555,718 SNPs present in the eMERGE data, as we observe that these SNPs tend to be well-imputed in Caucasian GWAS (Supplementary Table 4).

The importance of the heritability model

We begin by analyzing summary statistics from simulated phenotypes. Using the eMERGE data14,17 (25,875 individuals recorded for 4,555,718 SNPs), we generate 200 phenotypes each with 2,000 causal SNPs, $h^2_{SNP} = 0.5$, no confounding bias, and such that consecutive pairs (Phenotypes 1 & 2, 3 & 4, etc) have genetic correlation 0.5. For Phenotypes 1-100, we sample causal SNP effect sizes according to the GCTA Model, for Phenotypes 101-200, according to the LDAK Model. We perform single-SNP analysis for each phenotype, then analyze the resulting summary statistics using LDSC-Zero, LDSC, SumHer-Zero and SumHer-GC. Selected results are provided in Figure 1, with full results in Supplementary Figures 2 & 3.

First we examine results from analyzing each phenotype individually. Figure 1a shows that, as expected, accurate estimates of h^2_{SNP} are returned when phenotypes are analyzed assuming the matching heritability model (i.e., when GCTA phenotypes are analyzed using LDSC-Zero or LDSC and LDAK phenotypes are analyzed using SumHer-Zero or SumHer-GC), but that using a different heritability model can result in very poor estimates; SumHer-Zero tends to over-estimate h^2_{SNP} by about 20% when applied to GCTA phenotypes, while LDSC-Zero tends to under-estimate h^2_{SNP} by about 40% when applied to LDAK phenotypes. As shown in Supplementary Figure 2a, LDSC infers that there is only slight confounding when used on GCTA phenotypes (average intercept 1.008, s.d. 0.001), and therefore its estimates of h^2_{SNP} closely match those from LDSC-Zero. However, when used on LDAK phenotypes, LDSC wrongly infers that much of the causal signal is in fact confounding (average intercept 1.096, s.d. 0.001), and as a result, its estimates of h^2_{SNP} are on average about 60% lower than those from LDSC-Zero and about 75% lower than the true value.

Figure 1b reports the estimated enrichment of SNPs in conserved regions. As causal SNPs were picked at random from across the genome, the true enrichment is one. Again, we see that assuming the correct heritability model produces reliable estimates, but assuming the wrong model can result in misleading conclusions. In particular, we find that when LDSC-Zero or LDSC are used to analyze LDAK phenotypes, they infer that conserved regions are

over 2-fold enriched for heritability. We chose to focus on conserved regions as this was the category that, by applying LDSC to real data, Finucane *et al.*2 found to be most enriched. These simulations show that a substantial portion of the enrichment observed by Finucane *et al.* could be an artifact of misspecifying the heritability model.

Figure 1c compares results from analyzing consecutive pairs of phenotypes. Whereas LDSC-Zero and SumHer-Zero only produce accurate estimates of genetic correlation when applied to GCTA and LDAK phenotypes, respectively, we find that LDSC and SumHer-GC produce accurate estimates for both sets of phenotypes (Supplementary Fig. 2f shows why this is the case). Even so, highest precision is achieved when the correct heritability model is assumed; the s.d. of SumHer-GC estimates is on average about 50% higher than that of LDSC estimates when analyzing GCTA phenotypes, but about 50% lower when analyzing LDAK phenotypes.

Determining the most appropriate heritability model

With access to individual-level data, we can compare how well different heritability models fit real data using the likelihood from REML analysis.5 With only summary statistics, we can instead use the likelihood from SumHer. Supplementary Table 5 shows that across the 25 raw GWAS, the SumHer log likelihood is on average 17 nats higher under the LDAK Model than under the GCTA Model (for comparison, the average difference in REML log likelihood is 19 nats; Supplementary Table 6). Across the 24 summary GWAS, the SumHer log likelihood is on average 76 nats higher under the LDAK Model (Supplementary Table 7).

Rather than comparing based on likelihoods, an easier-to-visualize approach is to run SumHer using a hybrid heritability model containing both the GCTA and LDAK Models, and allow the data to decide the relative weighting of each. Specifically, we use Hybrid-Zero (or Hybrid-GC if the test statistics are confounded), with the focus on estimating p, the "proportion of LDAK", in Model (2). Figure 2a demonstrates proof of principle using our simulated phenotypes: we see that Hybrid-Zero correctly estimates $p \approx 0$ when applied to the GCTA phenotypes, $p \approx 1$ for the LDAK phenotypes, and $p \approx 0.5$ for "Hybrid Phenotypes" (created by summing Phenotypes 1 & 101, 2 & 102, etc.). For Figure 2b and Supplementary Table 5, we apply Hybrid-Zero to the 25 raw GWAS; the average estimate of p is 1.03 (s.d. 0.02), indicating that the data overwhelmingly support the LDAK Model over the GCTA Model. For Figure 2c and Supplementary Table 7, we apply Hybrid-GC to the 24 summary GWAS; the average estimate of p is 0.85 (s.d. 0.01), again strongly supporting the LDAK Model (in the Discussion, we consider why this estimate is lower than that for the 25 raw GWAS).

Population structure and relatedness

Supplementary Tables 2, 8 & 9 reports estimates of confounding bias, SNP heritability, enrichments and genetic correlations for the 25 raw GWAS. As a consequence of our strict quality control, SumHer-GC finds only slight confounding bias; its average estimate of the scaling factor C is 1.005 (s.d. 0.002). We now construct GWAS with substantial confounding. First, for each of the 13 WTCCC GWAS, we introduce population structure by

Europe PMC Funders Author Manuscripts

replacing 2000 of the controls (on average 54%) with 2000 individuals from POBI18 (People of the British Isles); this generates population structure because, although both WTCCC and POBI individuals were recruited from the UK, the latter predominately came from isolated, rural regions (Supplementary Fig. 4). Second, for each of the 12 eMERGE GWAS, we no longer filter individuals based on relatedness; this leads to the retention of approximately 1650 pairs of relatives (Supplementary Fig. 5). Supplementary Tables 10 & 11 confirm that in both cases, SumHer-GC now detects significant confounding; its average estimates of C are 1.022 (s.d. 0.003) and 1.020 (s.d. 0.003), respectively.

Genomic control and mixed-model association analysis

Although its use is declining, it remains that the majority of published GWAS have performed genomic control (divided test statistics by the GIF) at least once in their analyses.11 As the GIF tends to over-estimate confounding bias,19 genomic control will tend to result in deflated test statistics. Supplementary Figures 6 & 7 show that, if not accounted for, genomic control will result in under-estimation of h^2_{SNP} and inaccurate estimates of heritability enrichments, but that reliable estimates can be obtained by allowing for multiplicative inflation of test statistics. SumHer, like LDSC, is designed to be used with results from classical (least-squared) regression. However, with the development of software such as Fast-LMM, GCTA-LOCO and Bolt-LMM20–22 it is now common for GWAS to instead use mixed-model association analysis.23,24 Supplementary Figures 6, 8 & 9 show that when estimating h^2_{SNP} and enrichments, the impact of mixed-model analysis is similar to, albeit less severe than, that of genomic control. However, as with genomic control, reliable estimates can be obtained by allowing for multiplicative inflation of test statistics.

Estimates of confounding bias and SNP heritability

Table 1 reports estimates of confounding bias for each of the 24 summary GWAS. LDSC estimates the average intercept to be 1.042 (s.d. 0.002), indicating that the test statistics tend to be inflated. By contrast, SumHer-GC estimates the average scaling factor to be 0.929 (s.d. 0.003), indicating that the test statistics tend to be deflated. The widespread deflation observed by SumHer-GC is consistent with the fact that at least 15 of the 24 summary GWAS performed genomic control (Supplementary Table 3); in particular, we note that the four GWAS estimated to have lowest C (0.55 for body mass index, 0.68 for HDL cholesterol, 0.73 for LDL cholesterol and 0.70 for triglyceride levels), are all meta-analyses that used genomic control both before and after combining results across cohorts.25,26 Table 1 also reports the number of independent loci with $P < 5\times10\text{-}8$. Without adjustment, there are on average 86 loci per trait. If we adjust based on the results of LDSC (i.e., for each trait divide test statistics by the LDSC estimate of confounding bias), the average reduces to 62, whereas if we adjust based on the results of SumHer-GC, it increases to 118. For these counts, we defined two significant SNPs as dependent if they are within 1 cM and have $r^2_{jl} > 0.1$, but we find that results are very similar if instead we increase the window size to 3 cM or use $r^2_{jl} > 0.2$ (Supplementary Table 12).

Table 1 shows that across the 24 traits, estimates of h^2_{SNP} from SumHer-GC are on average 2.7 (s.d. 0.05) times higher than those from LDSC. As shown in Supplementary Table 13, this difference is primarily due to changing the heritability model, rather than changing the

confounding model; for example, were we to switch from the GCTA to LDAK Model but to continue using an additive model for confounding inflation, estimates would still be on average 2.3 (s.d. 0.04) times higher.

Estimates of functional enrichments and genetic correlations

Figure 3a and Supplementary Table 14 report estimates of enrichments for the 24 functional categories, averaged across the 24 traits. We see striking differences between the estimates from LDSC and SumHer-GC. For example, LDSC estimates of enrichments range from -1.1 to 9.4, whereas SumHer-GC estimates range from 0.78 to 2.0. The estimated enrichment of a category is its estimated share of h^2_{SNP} divided by its expected share under the assumed heritability model.27 Supplementary Table 14 shows that changing from the GCTA to LDAK model has a small impact on the expected share of each category (the denominator), but typically has a large impact on the estimated share (the numerator). Supplementary Figure 10 and Supplementary Tables 15 & 16 show that the large differences between LDSC and SumHer-GC estimates of enrichments remain if we calculate the former using the LDSC software1 (in this setting, LDSC is often referred to as stratified LDSC or S-LDSC) rather than using our implementation of LDSC in SumHer, and likewise if we vary the SNP sets (the authors of LDSC recommend that the reference panel contains as many SNPs as possible, but that only HapMap3 SNPs28 with MAF ≥ 0.05 are used when performing the regression1,2), or if we use the 75-part model proposed by Gazal et al.29

Based on the results from SumHer-GC, we conclude that conserved regions30,31 and transcription start sites32 are most enriched for heritability; they are estimated to contribute 7.1% (s.d. 0.2) and 3.6% (s.d. 0.2) of h^2_{SNP}, respectively, 1.95 (s.d. 0.07) and 1.97 (s.d. 0.09) times higher than their expected contributions under the LDAK Model. Repressed region are the only category significantly depleted; although they are estimated to contribute 35% (s.d. 0.5) of h^2_{SNP}, this is only 0.78 (s.d. 0.01) times their expected contribution.

Supplementary Figure 11 and Supplementary Table 17 provide estimates of genetic correlation for the 276 pairs of traits. As predicted by our simulation study, there is strong concordance between estimates from LDSC and SumHer-GC, but the SumHer-GC estimates are more precise; for example, across the 38 pairs of traits that both methods find to be significantly correlated ($P < 0.05/276$), the s.d. of SumHer-GC estimates is on average about 20% lower than that of LDSC estimates.

Improving polygenic risk scores

Figure 3b shows that there is strong concordance between the average estimates of enrichments obtained from the 9 binary traits and those from the 15 quantitative traits. This suggests broad similarities between the genetic architectures of different traits, which in turn implies that it should be possible to use information from existing GWAS to improve the efficiency of future analyses. As a demonstration, we consider prediction using polygenic risk scores (PRS). We focus on body mass index, height, HDL & LDL cholesterol and triglyceride levels, as for these five traits we can train prediction models using the 24 summary GWAS, then measure how well these perform using the (independent) eMERGE data.14

Europe PMC Funders Author Manuscripts

To construct a PRS, we need estimates of SNP effect sizes. The most common approach is to use estimates from single-SNP analysis33 ("Classical PRS"). However, in Online Methods, we explain how, given a heritability model, we can obtain a prior distribution for v^2_j, the heritability tagged by SNP j, then calculate a "Bayesian PRS" using the posterior mean effect sizes. For each trait, we construct four Bayesian PRS corresponding to four heritability models. First we use the GCTA heritability model ($q_j = 1$). Next we use the "Enriched GCTA Model", obtained by scaling the q_j based on the LDSC estimates of enrichments (e.g., if a category was estimated to have 2-fold enrichment, then the SNPs it contains would have average $q_j = 2$). We similarly construct PRS based on the LDAK Model, then the "Enriched LDAK Model", where the q_j are scaled according to the SumHer-GC estimates of enrichments.

Figure 4a compares the four prior distributions for v^2_j. We see that the two priors derived from the GCTA Model are more diffuse than the two from the LDAK Model. As explained in the Online Methods, this is because the GCTA Model predicts that v^2_j scales with local levels of LD, which vary considerably across the genome, whereas the LDAK Model predicts that v^2_j scales with local MAFs, which vary less. Figure 4b and Supplementary Table 18 report the performance of each PRS, measured as correlation between predicted and observed phenotypes for the eMERGE individuals. Averaged across the five traits, the Bayesian PRS constructed from the GCTA Model performs 0.1% (s.d. 1.6) worse than the Classical PRS (i.e., no significant difference). However, the Bayesian PRS constructed from the Enriched GCTA, LDAK and Enriched LDAK Models perform, respectively, 5.2%, 4.7% and 6.1% better (s.d.s all 1.6), each of which is a significant improvement ($P < 0.05/4$) over the Classical PRS.

Discussion

We have presented SumHer, software for estimating confounding bias, SNP heritability, enrichments of heritability and genetic correlations from GWAS results. While the aims of SumHer are the same as those of LDSC, the key difference is that SumHer allows the user to specify the heritability model. If SumHer is run using the GCTA Model, its estimates will match those from LDSC. However, we instead recommend using the LDAK Model, which we have shown is better supported by real data. We have analyzed GWAS results for tens of traits, showing that the impact of using an improved heritability model is often substantial, and overall provides a very different description of the genetic architecture of complex traits than has to date been obtained from LDSC analyses.

The heritability model matters

The heritability model describes the prior assumptions regarding how heritability is distributed across the genome. We began by showing that estimates of confounding bias, h^2_{SNP} and enrichments are sensitive to the assumed heritability model, and that using an unrealistic heritability model can result in misleading conclusions. Next we evaluated the GCTA and LDAK heritability models on real data. Not only did we demonstrate that the LDAK Model is substantially more realistic than the GCTA Model, but also our prediction analysis showed that the LDAK Model is sufficiently realistic to be useful. Nonetheless, it is

important to recognize that the LDAK heritability model is not perfect, and as such SumHer provides tools for testing and comparing different heritability models on large-scale GWAS data.

Hybrid heritability models

As well as comparing the GCTA and LDAK Models based on likelihood, we also ran SumHer using a hybrid heritability model where the fractions 1-p and p indicate the proportions of GCTA and LDAK, respectively; across the 25 raw GWAS, the average estimate of p was 1.03 (s.d. 0.02), while across the 24 summary GWAS, the average estimate of p was 0.85 (s.d. 0.01). Although both estimates of p are greater than 0.5, consistent with the LDAK Model being more realistic than the GCTA Model, the latter suggests the hybrid model might improve on the LDAK Model. We recommend using only the LDAK Model for two reasons. Firstly, estimates from the LDAK Model tend to be more precise than those from the hybrid model. This is particularly true when estimating enrichments; for example, Supplementary Table 19 shows that across the 24 summary GWAS, the s.d. of SumHer-GC estimates is on average about 70% lower than that of Hybrid-GC estimates. Secondly, Supplementary Figure 12 shows that introducing poorly-genotyped SNPs makes traits appear more GCTA-like, indicating that the hybrid model is more sensitive to genotyping errors. In any case, Supplementary Figure 13 shows that even if we had instead preferred results from the hybrid model, it would not have affected our overall conclusions.

Confounding bias in GWAS

In addition to examining the choice of heritability model, we also considered how to model inflation of test statistics due to confounding. Whereas LDSC uses an additive model for confounding inflation, we prefer a multiplicative model, due to the fact that genomic control (which scales test statistics multiplicatively) has been widely used by published GWAS (we provide further support for our choice in Supplementary Fig. 14 and Supplementary Table 13). We recognize that as genomic control becomes less common, it will be necessary to check whether the multiplicative model remains preferable to the additive model.

Implications for complex trait genetics

Our work has three main conclusions. **(i)** Large-scale GWAS have tended to over-correct for confounding, which substantially reduced the discovery of genome-wide significant associations. Although SumHer estimates of confounding bias can be used to adjust test statistics from a confounded GWAS (as we did for Table 1, in order to estimate the number of genome-wide significant loci missed by the 24 summary GWAS), adjusted test statistics will not be as accurate as those from a GWAS without confounding. Therefore, we would prefer that SumHer is used during the primary analysis; if SumHer finds substantial confounding bias, this indicates that the analysis should be repeated with improved quality control. **(ii)** LDSC tends to substantially under-estimate h^2_{SNP}. Accurate estimates of h^2_{SNP} are important not only because they improve our understanding of genetic architecture,16 but also because they are now being incorporated in software for analyzing complex traits (e.g., the mixed-model association software Bolt-LMM22 and the prediction software LDPred34). The efficiency of these analyses can therefore be improved by incorporating

Europe PMC Funders Author Manuscripts

SumHer estimates of h^2_{SNP}. **(iii)** The most striking differences between LDSC and SumHer are observed when estimating enrichments of heritability in functional regions. Whereas analyses using LDSC have found that heritability is concentrated in specific genomic regions, we have instead shown that heritability is spread more evenly across the genome. Although the realization that complex traits are even more complicated than previously thought is daunting, it is only by properly understanding their complexity that we can develop more efficient tools for analyzing genetic data.

Methods

Estimating SNP heritability

Suppose that we have summary statistics from a GWAS on n individuals and m SNPs; let S_j denote the $\chi^2(1)$ test statistic from regressing the phenotype on X_j, the vector of additively-coded genotypes for SNP j, and let $n_j \leq n$ denote the number of individuals used in this regression (note that in the main text, for simplicity, we assumed $n_j = n$). If S_j was obtained using classical (i.e., least-squares) linear regression, then19

$$E[S_j] \approx 1 + n_j v_j^2 \quad \text{with} \quad v_j^2 = h_j^2 + \sum_{l \neq j} \mathrm{Cor}(X_j, X_l)^2 h_l^2, \quad (3)$$

where v^2_j is the total amount of heritability tagged by SNP j. In the main text, we referred to h^2_j as the heritability "directly contributed" by SNP j, to emphasize that while a causal signal can contribute to multiple v^2_j (i.e., be tagged by multiple SNPs), it can only contribute to one h^2_j. More formally, the h^2_j represent a partitioning of h^2_{SNP}, the total heritability tagged by the m SNPs genotyped by the GWAS; this definition recognises that an ungenotyped causal variant can contribute to h^2_{SNP}, provided it is tagged by one or more genotyped SNPs (in which case its heritability will be shared across the h^2_j of the tagging SNPs, even though none of these "directly contribute" its heritability). If there is no population structure or familial relatedness, then $\mathrm{Cor}(X_j,X_l)^2$ will be negligible for distant SNPs, while for local SNPs an unbiased estimate of $\mathrm{Cor}(X_j,X_l)^2$ is1

$$r^2_{jl} = \mathrm{Cor}(X'_j, X'_l)^2 - \frac{1 - \mathrm{Cor}(X'_j, X'_l)^2}{n' - 2}$$

where X'_j is the vector of SNP j genotypes for the n' individuals in a reference panel (for accurate estimates of r^2_{jl}, the individuals in the reference panel should have similar ancestry to those used in the GWAS). Therefore, in place of Equation (3) we use

$$E[S_j] \approx 1 + n_j v_j^2 \quad \text{with} \quad v^2 = h_j^2 + \sum_{l \in N_j} r^2_{jl} h^2_l, \quad (4)$$

where the set N_j indexes those SNPs "near" SNP j (a working definition of near is within 1 cM; Supplementary Fig. 15). Finally, given a heritability model of the form $E[h^2_j] \propto q_j$, we replace h^2_j by its expected value $q_j\, h^2_{SNP} / Q$, where $Q = \sum q_j$, resulting in

$$E[S_j] \approx 1 + u_j h^2_{SNP} \quad \text{with} \quad u_j = n_j(q_j + \sum_{l \in N_j} q_l r^2_{jl})/Q, \quad (5)$$

which allows us to estimate h^2_{SNP} by regressing $(S_j - 1)$ on u_j. To account for correlated datapoints and heteroscedasticity we use weighted least-squares regression. Specifically, if D is a diagonal matrix whose non-zero entries are the regression weights, then the estimate of h^2_{SNP} would be $(u^T D u)^{-1} u^T D (S - 1)$, where u and S are vectors containing the m values for u_j and S_j, respectively. Following LDSC,1 we use $1 / D_{jj} = (\sum_{l \in Nj} r^2_{jl})(1 + u_j h^2_{SNP})$. Since h^2_{SNP} is unknown, we proceed iteratively starting at $h^2_{SNP} = 0$, then successively updating D_{jj} and h^2_{SNP} until convergence. Again following LDSC, we estimate standard errors via block jackknifing (by default we use 200 blocks).1

Comparing heritability models

The (weighted) natural log likelihood is

$$L(S|h^2_{SNP}, D) = \frac{-d}{2}(\log(2\pi\lambda/d) + 1) \quad \text{with} \quad d = \sum_j D_{jj} \quad \text{and} \quad \lambda = \sum_j D_{jj}(S_j - 1 - u_j h^2_{SNP})^2$$

To evaluate $L(S \mid h^2_{SNP}, D)$ requires values for h^2_{SNP} and D. While the natural choice is to use the values after the final iteration, in order to compare different heritability models based on $L(S \mid h^2_{SNP}, D)$, we must use the same D for each (else models leading to lower D_{jj} would have an unfair advantage). Therefore, SumHer reports $L(S \mid \widehat{h^2_{SNP}}, D^0)$, where $\widehat{h^2_{SNP}}$ is the final estimate of h^2_{SNP}, but D^0 is the initial weight matrix (obtained by setting $1 / D_{jj} = \sum_{l \in Nj} r^2_{jl}$); Supplementary Tables 5 & 6 show that for the 25 raw GWAS, comparisons of heritability models based on $L(S \mid \widehat{h^2_{SNP}}, D^0)$ align closely with those based on REML likelihood.

Estimating confounding bias

When assuming that confounding inflates test statistics multiplicatively, we replace Equation (5) with $E[S_j] \approx C(1 + u_j h^2_{SNP})$, then estimate C and h^2_{SNP} by jointly regressing S_j on 1 and u_j. To instead copy LDSC and assume confounding causes additive inflation, we replace Equation (5) with $E[Sj] \approx 1 + n_j a + u_j h^2_{SNP}$, then estimate a and h^2_{SNP} by jointly regressing $(S_j - 1)$ on n_j and u_j (note that in the main text we assumed $n_j = n$, allowing us to replace $n_j a$ by A).

Estimating enrichments

Suppose we have K categories; let $I_{jk} \in \{0,1\}$ indicate whether SNP j belongs to Category k. We wish to estimate $h^2_{Cat\ k} = \sum_j I_{jk} h^2_j$, the heritability contributed by SNPs in Category k. We now use the heritability model

Europe PMC Funders Author Manuscripts

$$E[h_j^2] = q_j \sum_k I_{jk} h^2_{partk} / Q_k \quad \text{with} \quad Q_k = \sum_j I_{jk} q_j \quad (6)$$

where $q_j\ I_{jk}\ h^2_{Part\ k} / Q_k$ represents the contribution of Category k to the expected heritability of SNP j, and $h^2_{SNP} = h^2_{Part\ 1} + \ldots + h^2_{Part\ K}$. We consider two scenarios: in Scenario 1, the K categories partition the genome ($\sum_k I_{jk} = 1$); in Scenario 2, the first K-1 categories correspond to annotations, and the Kth is the base category containing all SNPs ($I_{jK} = 1$, $Q_K = m$). Using Equation (6) ensures that when there is no enrichment, $E[h^2_j] = q_j\ h^2_{SNP} / Q$ for both scenarios. To appreciate why, consider that for Scenario 1, $h^2_{Part\ k} = h^2_{Cat\ k}$, so that when no categories are enriched, $h^2_{Part\ k} = Q_k\ h^2_{SNP} / Q$; for Scenario 2, $h^2_{Part\ 1}, \ldots, h^2_{Part\ K-1}$ indicate how much each annotation increases the SNP heritability from $h^2_{Part\ K}$, so that when there is no enrichment, $h^2_{Part\ 1} = \ldots = h^2_{Part\ K-1} = 0$ and $h^2_{Part\ K} = h^2_{SNP}$. Now when we replace h^2_j in Equation (4) by its expected value, we obtain

$$E[S_j] \approx 1 + \sum_k u_{jk} h^2_{part\ k} \qquad \text{with} \quad u_{jk} = n_j (q_j I_{jk} + \sum_{l \in N_j} q_j I_{lk} r^2_{jl}) / Q_k$$

and therefore we can estimate $h^2_{Part\ 1}, \ldots, h^2_{Part\ K}$ by jointly regressing (S_j - 1) on $u_{j1}, \ldots, u_{jK}$. Given these, our estimate of $h^2_{Cat\ k}$ is $\sum_{j,k'} (I_{jk}\ q_j\ I_{jk'}\ h^2_{Part\ k'} / Q_{k'})$, which we then divide by $Q_k\ h^2_{SNP} / Q$ to get an estimate of the enrichment of Category k.

Estimating genetic correlation

Suppose we have summary statistics from two GWAS. Instead of $\chi^2(1)$ test statistics, we now use (signed) Z-statistics. Let Z_{Aj} and Z_{Bj} denote the two Z-statistics for SNP j, computed using n_{Aj} and n_{Bj} individuals, respectively, of which n_{Cj} were common to both GWAS (if the two GWAS were independent, $n_{Cj} = 0$). We assume

$$E[Z_{A_j} Z_{B_j}] \approx \frac{c_{AB} n_{S_j}}{\sqrt{n_{A_j} n_{B_j}}} + u'_j h^2_{AB} \quad \text{with} \quad u'_j = \sqrt{n_{A_j} n_{B_j}} (q_j + \sum_{i \in N_j} q_l r^2_{jl}) / Q$$

where c_{AB} is the phenotypic correlation between the two traits and h^2_{AB} is their genetic covariance. This equation matches that used by LDSC,3 except that we have replaced r^2_{jl} / m by $q_j\ r^2_{jl} / Q$. By regressing $Z_{Aj} Z_{Bj}$ on u'_j, we obtain an estimate of h^2_{AB}, which we then divide by estimates of $\sqrt{h^2_{SNP}}$ for each trait to get an estimate of their genetic correlation.

Impact of changing the confounding model

Note that when per-SNP sample sizes are constant, $n_j = n$, so we can write $1 + n_j\ a + u_j\ h^2_{SNP}$ as $C\ (1 + u_j\ h^2_{SNP} / C)$, where $C = 1 + n\ a = 1 + A$. Therefore, given the heritability model (which determines the u_j), the estimate of confounding bias will be the same whether we assume additive inflation of test statistics (and estimate 1+A) or assume multiplicative inflation (and estimate C), but estimates of h^2_{SNP} from the additive model will be C times

Europe PMC Funders Author Manuscripts

those from the multiplicative model (estimates of enrichments and genetic correlations will be unaffected, as these are ratios of heritabilities so the factor C cancels).

Reference panel

Our primary reference panel is 404 non-Finnish, European individuals from the 1000 Genomes Project.12 In Supplementary Figure 16, we repeat our analyses of the 24 summary GWAS using instead 8,850 unrelated Caucasian individuals from the Health & Retirement Study. The results are almost identical, indicating that despite its small size, it suffices to construct reference panels from the 1000 Genomes Project data.

Regions of extreme LD and large-effect loci

Estimates can be unduly affected by regions of extreme LD and by SNPs with disproportionately large effect size.1,4,5 Therefore, for all analyses, we exclude SNPs within the MHC (Chromosome 6: 25-34 Mb), as well as SNPs which individually explain >1% of phenotypic variation ($S_j > n_j / 99$), and SNPs in LD with these (within 1 cM and $r^2_{jl} > 0.1$); the latter resulted in the exclusion of SNPs for 10 of the 25 raw GWAS and for 4 of the 24 summary GWAS (Supplementary Table 20). Excluding large-effect loci will result in under-estimation of h^2_{SNP}; this has little consequence for our analyses, as we are primarily interested in comparing models, but if preferred can be avoided by re-introducing the large-effect SNPs as fixed effects.5

Binary phenotypes

So far, we have implicitly assumed the trait is quantitative. If instead it is binary (i.e., for case-control GWAS), there are two considerations. Firstly, heritability estimates now correspond to the observed scale; SumHer will convert them to the liability scale if provided with the prevalence and case-control ratio.51,52 Secondly, the p-values may have come from logistic regression instead of linear regression; Supplementary Figure 17 shows that because linear and logistic p-values closely match for SNPs with small or moderate effect, this has limited impact.

Polygenic risk scoring

To predict phenotypes, we construct PRS of the form $\sum_j \beta_j X'_j$, where the vector X'_j contains standardized genotypes for SNP j (obtained by centering X_j, then scaling to have variance 1). Without loss of generality, we assume phenotypes have also been standardized, in which case the estimate of β_j from classical linear regression is ρ_j, the correlation observed between SNP j and the phenotype, and has variance $(1 - \rho^2_j) / n_j$. We can infer ρ_j from the Wald test Z-statistic $p_j/\sqrt{(1-p_j^2)/n_j}$. For the Classical PRS, our estimate of β_j is ρ_j. For the Bayesian PRS, we use the prior distribution $\beta_j \sim \mathbb{N}(0, \sigma^2_j)$, where $\sigma^2_j = (q_j + \sum_{l \in Nj} q_l r^2_{jl}) h^2_{SNP} / Q$; this choice is motivated by recognizing that for the "true PRS", $\beta^2_j = v^2_j$, the heritability tagged by SNP j. We then estimate β_j by its posterior mean, a shrunken version of the classical estimate. To calculate this, we approximate the likelihood by $\mathbb{N}(\rho_j, (1 - \rho^2_j) / n_j)$, then the posterior mean equals $\rho_j \sigma^2_j / (\sigma^2_j + (1 - \rho^2_j) / n_j)$.53,54 There are two technical issues. Firstly, to construct each Bayesian PRS requires a value for h^2_{SNP}, so we used the corresponding estimates from LDSC-Zero and SumHer-Zero; Supplementary Table 18

confirms that predictions are approximately the same if we instead use the estimates from LDSC and SumHer-GC, or agnostically set $h^2_{SNP} = 0.5$. Secondly, for the Bayesian PRS, we performed clumping (identified pairs of SNPs within 1 cM with $r^2_{jl} > 0.5$, then discarded the least significant), whereas for the Classical PRS we did not; this was because we found that clumping benefited all Bayesian PRS, but did not benefit the Classical PRS.

Choosing a heritability model

Although the heritability model is defined in terms of h^2_j, as is clear from Equation (4), its fit depends only on how well it models $v^2_j = h^2_j + \sum_{l \in Nj} r^2_{jl} h^2_l$. Therefore, when specifying q_j, the focus should be on accurately describing how v^2_j varies across the genome. LDSC assumes the GCTA Model,5,7 ($q_j = 1$), which implies that v^2_j correlates with local levels of LD. Despite its widespread use in statistical genetics (e.g., it is implicitly assumed by any regression method where SNPs are standardized and the same penalty function / prior distribution is applied to each), the GCTA Model poorly reflects real data.5 Even for traits where a correlation is observed between v^2_j and S_j (those with $p < 1$ in Figures 2b & 2c), the correlation is substantially weaker than predicted by the GCTA Model (hence $p >> 0$), and partly a consequence of including lower-certainty SNPs in the GWAS (Supplementary Fig. 12). We recommend using the LDAK Model. Originally, this took the form $q_j = w_j$, where the weightings w_j are calculated based on the assumption that $E[v^2_j]$ is constant.6 We recently updated it to $q_j = w_j [f_j(1 - f_j)]^{0.75} r_j$, reflecting that v^2_j tends to depend on MAF, f_j, and genotype certainty, r_j (we did not use the latter as either $r_j \approx 1$ or was unknown).5 The exponent 0.75 was chosen as this value performed well in a previous analysis of 42 human traits. In Supplementary Figure 18, we confirm 0.75 also performs well for both the 25 raw and 24 summary GWAS.

Quality control

To prepare the 13 WTCCC GWAS, we used our previously-described protocol.5 In summary, after excluding apparent population outliers, individuals with extreme missingness or heterozygosity and SNPs with MAF < 0.01, call rate < 0.95 or Hardy-Weinberg $P < 1 \times 10^{-6}$, we phased using SHAPEIT, then imputed using IMPUTE2 with the 1000 Genomes Project Phase 3 (2014) reference panel.12,55,56 When merging case and control datasets, we converted genotype probabilities to hard calls using a threshold of 0.95, then retained only autosomal SNPs that in all cohorts had MAF ≥ 0.01 and info score ≥ 0.99. Finally, we thinned individuals, so no pair remained with allelic correlation57 > 0.05. For the 12 eMERGE GWAS, data were provided post-imputation; this was also performed using SHAPEIT and IMPUTE2, but used the 1000 Genomes Project Phase 2 reference panel.17 We converted genotype probabilities to hard calls using a certainty threshold of 0.95, then retained only biallelic SNPs with MAF ≥ 0.01, call rate ≥ 0.95, info score ≥ 0.95 and whose genomic positions matched those in the 1000 Genomes Project. Finally, we excluded individuals whose genotypes were inconsistent with those of non-Finnish Europeans from the 1000 Genomes Project (Supplementary Fig. 19), and those whose ethnicity was reported as “Hispanic or Latino”, then filtered until no pair remained with allelic correlation57 > 0.05 (which left 25,875 individuals). For the 24 summary GWAS, quality control varied by study (see Table 1 for references), but was in general much less strict than we would have performed had we access to the individual-level data16 (for example, the most common info

score thresholds were 0.3 and 0.5). Info scores were only available for Crohn's Disease, inflammatory bowel disease, schizophrenia and ulcerative colitis; for these traits we restricted to SNPs with score ≥ 0.95, for the remainder, we restricted to SNPs with info score ≥ 0.95 in the eMERGE data (Supplementary Table 4).

Association analysis

For the 25 raw GWAS, we performed the association analysis using linear regression (regardless of whether the trait was quantitative or binary), including as covariates sex and ten principal components (five derived from the reference panel, five from the 1000 Genomes Project). The 24 summary GWAS varied in whether they used classical or mixed-model regression, and whether they performed a meta or mega analysis (Supplementary Table 3).

Software

SumHer is implemented in Version 5.0 of the LDAK software. In Supplementary Figure 20 we confirm that LDSC estimates computed using SumHer match those computed using Version 1.0.0 of the LDSC Software.1

Run times

To run SumHer using the LDAK Model requires three steps: the first is to calculate SNP weightings; the second is to compute a "tagfile", which contains $q_j + \sum_{l \in Nj} q_l r^2_{jl}$ for each SNP; the third is to perform the regression. When using the 404 non-Finnish Europeans from the 1000 Genomes Project, the first step takes a few hours, the second and third take minutes. When running SumHer using the GCTA Model, it is not necessary to calculate the SNP weightings, but then computing the tagfile will generally take a few hours.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank A Price, H Finucane, P O'Reilly and M Speed for helpful discussions. Access to Wellcome Trust Case Control Consortium data was authorized as work related to the project "Genome-wide association study of susceptibility and clinical phenotypes in epilepsy", access to eMERGE Network data was granted under dbGaP Project 14422, "Comprehensive testing of SNP-based prediction models", while access to the Health and Retirement Study was granted under dbGaP Project 15139, "Developing summary-statistic tools for analysing genetic association study data". D.S. is funded by the UK Medical Research Council under grant MR/L012561/1, by the European Unions Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie grant agreement number 754513, by Aarhus University Research Foundation (AUFF) and the Independent Research Fund Denmark under Project 7025-00094B. The eMERGE Network was initiated and funded by NHGRI through the following grants: U01HG006828 (Cincinnati Childrens Hospital Medical Center/Boston Childrens Hospital); U01HG006830 (Childrens Hospital of Philadelphia); U01HG006389 (Essentia Institute of Rural Health, Marshfield Clinic Research Foundation and Pennsylvania State University); U01HG006382 (Geisinger Clinic); U01HG006375 (Group Health Cooperative); U01HG006379 (Mayo Clinic); U01HG006380 (Icahn School of Medicine at Mount Sinai); U01HG006388 (Northwestern University); U01HG006378 (Vanderbilt University Medical Center); and U01HG006385 (Vanderbilt University Medical Center serving as the Coordinating Center). The Health and Retirement Study genetic data is sponsored by the National Institute on Aging (grant numbers U01AG009740, RC2AG036495, and RC4AG039029) and was conducted by the University of Michigan. Analyses were performed with the use of the UCL Computer Science Cluster and the help of the CS Technical Support Group, as well as the use of the UCL Legion High-Performance Computing Facility (Legion@UCL) and associated support services.

Europe PMC Funders Author Manuscripts

References

1. Bulik-Sullivan B, et al. LD Score Regression Distinguishes Confounding from Polygenicity in Genome-Wide Association Studies. Nat Genet. 2014; 47:291–295.
2. Finucane H, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat Genet. 2015; 47:1228–1235. [PubMed: 26414678]
3. Bulik-Sullivan B, et al. An atlas of genetic correlations across human diseases and traits. Nat Genet. 2015; 47:1236–1241. [PubMed: 26414676]
4. Zheng J, et al. LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. Bioinformatics. 2016; 33:272–279. [PubMed: 27663502]
5. Speed D, Cai N, Johnson MR, Nejentsev S, Balding DJ. Reevaluation of SNP heritability in complex human traits. Nat Genet. 2017; 49:986–992. [PubMed: 28530675]
6. Speed D, Hemani G, Johnson MR, Balding DJ. Improved heritability estimation from genome-wide SNPs. Am J Hum Genet. 2012; 91:1011–21. [PubMed: 23217325]
7. Yang J, et al. Common SNPs explain a large proportion of the heritability for human height. Nat Genet. 2010; 42:565–569. [PubMed: 20562875]
8. Yang J, et al. Genomic partitioning of genetic variation for complex traits using common SNPs. Nat Genet. 2011; 43:519–525. [PubMed: 21552263]
9. Lee SH, Yang J, Goddard ME, Visscher PM, Wray NR. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. Bioinformatics. 2012; 28:2540–2. [PubMed: 22843982]
10. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: A tool for genome-wide complex trait analysis. Am J Hum Genet. 2011; 88:76–82. [PubMed: 21167468]
11. Devlin B, Roeder K. Genomic control for association studies. Biometrics. 1999; 55:997–1004. [PubMed: 11315092]
12. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature. 2010; 467:1061–1073. [PubMed: 20981092]
13. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007; 447:661–678. [PubMed: 17554300]
14. Gottesman O, et al. The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. Genet Med. 2013; 15:761. [PubMed: 23743551]
15. Pasaniuc B, Price AL. Dissecting the genetics of complex traits using summary association statistics. Nat Rev Genet. 2017; 18:117–127. [PubMed: 27840428]
16. Speed D, et al. Describing the genetic architecture of epilepsy through heritability analysis. Brain. 2014; 137:2680–2689. [PubMed: 25063994]
17. Verma S, et al. Imputation and quality control steps for combining multiple genome-wide datasets. Front Genet. 2015; 5:370.
18. Leslie S, et al. The fine-scale genetic structure of the British population. Nature. 2015; 519:309–314. [PubMed: 25788095]
19. Yang J, et al. Genomic inflation factors under polygenic inheritance. Eur J Hum Genet. 2011; 19:807–812. [PubMed: 21407268]
20. Lippert C, et al. FaST linear mixed models for genome-wide association studies. Nat Methods. 2011; 8:833–835. [PubMed: 21892150]
21. Yang J, Zaitlen N, Goddard M, Visscher P, Price A. Advantages and pitfalls in the application of mixed-model association methods. Nat Genet. 2014; 46:100–106. [PubMed: 24473328]
22. Loh P, et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nat Genet. 2015; 47:284–290. [PubMed: 25642633]
23. Yu J, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet. 2006; 38:203–208. [PubMed: 16380716]
24. Sawcer S, et al. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature. 2011; 476:214–219. [PubMed: 21833088]

25. Locke A, et al. Genetic studies of body mass index yield new insights for obesity biology. Nature. 2015; 518:197–206. [PubMed: 25673413]
26. Global Lipids Genetics Consortium. Discovery and refinement of loci associated with lipid levels. Nat Genet. 2013; 45:1274–1283. [PubMed: 24097068]
27. Finucane H, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat Genet. 2015; 47:1228–1235. [PubMed: 26414678]
28. The International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. Nature. 2010; 467:52–58. [PubMed: 20811451]
29. Gazal S, et al. Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. Nat Genet. 2017; 49:1421–1427. [PubMed: 28892061]
30. Lindblad-Toh K, et al. A high-resolution map of human evolutionary constraint using 29 mammals. Nature. 2011; 478:476–482. [PubMed: 21993624]
31. Ward L, Kellis M. Evidence of abundant purifying selection in humans for recently acquired regulatory functions. Science (80-.). 2012; 337:1675–1678.
32. Hoffman M, et al. Integrative annotation of chromatin elements from ENCODE data. Nucleic Acids Res. 2013; 41:827–841. [PubMed: 23221638]
33. Euesden J, Lewis C, O'Reilly P. PRSice: polygenic risk score software. Bioinformatics. 2015; 31:1466–1468. [PubMed: 25550326]
34. Vilhjálmsson B, et al. Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. Am J Hum Genet. 2015; 97:576–592. [PubMed: 26430803]
35. Scott R, et al. An Expanded Genome-Wide Association Study of Type 2 Diabetes in Europeans. Diabetes. 2017; 66:2888–2902. [PubMed: 28566273]
36. Liu J, et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nat Genet. 2015; 47:979–986. [PubMed: 26192919]
37. Zheng H, et al. Whole-genome sequencing identifies EN1 as a determinant of bone density and fracture. Nature. 2015; 526:112–117. [PubMed: 26367794]
38. Okbay A, et al. Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. Nat Genet. 2016; 48:626–633.
39. Manning A, et al. A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. Nat Genet. 2012; 44:659–669. [PubMed: 22581228]
40. Soranzo N, et al. Common variants at 10 genomic loci influence hemoglobin A_1(C) levels via glycemic and nonglycemic pathway. Diabetes. 2010; 59:3229–3239. [PubMed: 20858683]
41. Wood A, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. Nat Genet. 2014; 46:1173–1186. [PubMed: 25282103]
42. Perry J, et al. Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche. Nature. 2014; 514:92–97. [PubMed: 25231870]
43. Day F, et al. Large-scale genomic analyses link reproductive aging to hypothalamic signaling, breast cancer susceptibility and BRCA1-mediated DNA repair. Nat Genet. 2015; 47:1294–1303. [PubMed: 26414677]
44. Shungin D, et al. New genetic loci link adipose and insulin biology to body fat distribution. Nat Genet. 2015; 518:187–196.
45. Okbay A, et al. Genome-wide association study identifies 74 loci associated with educational attainment. Nature. 2016; 533:539–542. [PubMed: 27225129]
46. Lambert J, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat Genet. 2013; 45:1452–1458. [PubMed: 24162737]
47. Schunkert H, et al. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. Nat Genet. 2011; 43:333–338. [PubMed: 21378990]
48. The Tobacco and Genetics Consortium. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. Nat Genet. 2010; 42:441–447. [PubMed: 20418890]
49. Okada Y, et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature. 2014; 506:376–381. [PubMed: 24390342]

50. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. Nature. 2014; 511:421–427. [PubMed: 25056061]
51. Dempster E, Lerner I. Heritability of threshold characters. Genetics. 1950; 35:212–236. [PubMed: 17247344]
52. Lee S, Wray N, Goddard M, Visscher P. Estimating missing heritability for disease from genome-wide association studies. Am J Hum Genet. 2011; 88:294–305. [PubMed: 21376301]
53. Wakefield J. Bayes factors for genome-wide association studies: comparison with p-values. Genet Epidemiol. 2009; 33:79–86. [PubMed: 18642345]
54. Pickrell J. Joint Analysis of Functional Genomic Data and Genome-wide Association Studies of 18 Human Traits. Am J Hum Genet. 2014; 94:559–573. [PubMed: 24702953]
55. Delaneau O, Zagury J, Marchini J. Improved whole-chromosome phasing for disease and population genetic studies. Nat Methods. 2013; 10:5–6. [PubMed: 23269371]
56. Howie B, Marchini J, Stephens M. Genotype imputation with thousands of genomes. G3. 2011; 1:457–470. [PubMed: 22384356]
57. Astle W, Balding DJ. Population structure and cryptic relatedness in genetic association studies. Stat Sci. 2009; 24:451–471.

Europe PMC Funders Author Manuscripts

Reporting summary

Further information on research design is available in the Life Sciences Reporting Summary linked to this article.

Europe PMC Funders Author Manuscripts

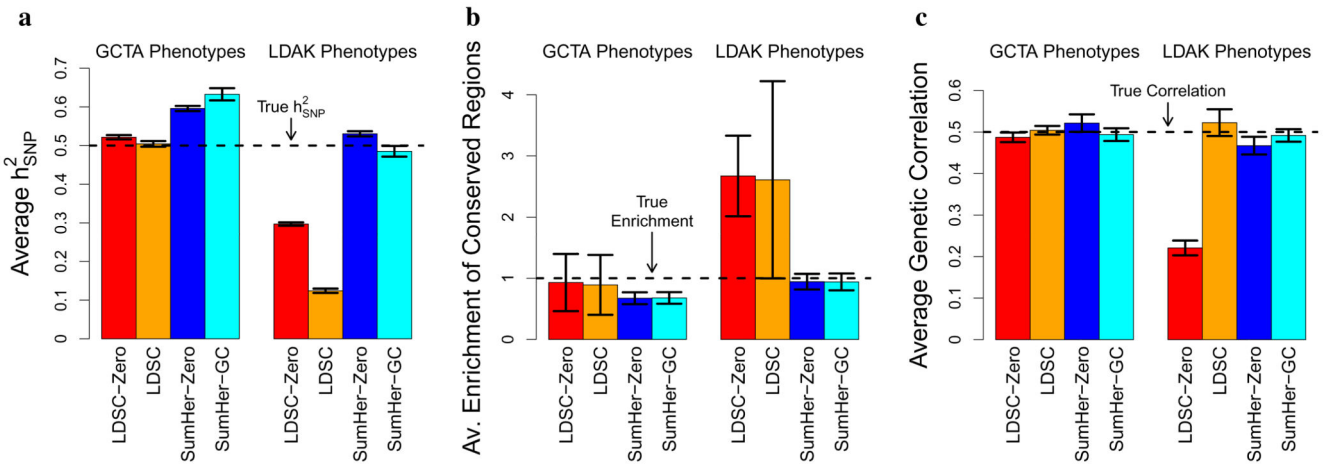

Figure 1. Importance of the heritability model.

We generated 100 phenotypes assuming the GCTA heritability model, 100 assuming the LDAK heritability model, then analyzed each using LDSC-Zero, LDSC, SumHer-Zero and SumHer-GC (see main text for details of the heritability models and methods). **a**, Average estimates of h^2_{SNP} (true h^2_{SNP} is 0.5). **b**, Average estimates of the enrichment of heritability in conserved regions (true enrichment is 1). **c**, Average estimates of genetic correlation between pairs of phenotypes (true correlation is 0.5). In all plots, vertical line segments mark 95% confidence intervals for the average estimates.

Europe PMC Funders Author Manuscripts

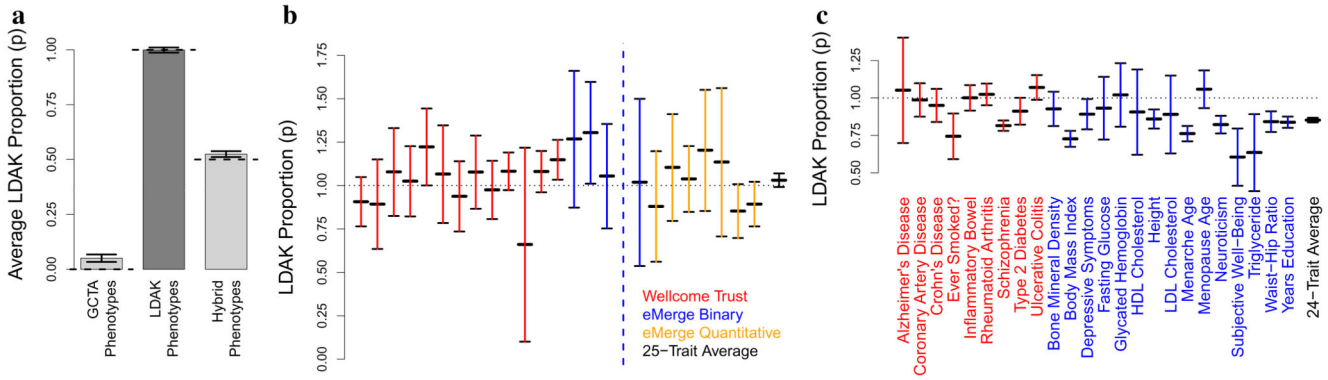

Figure 2. Comparing the GCTA and LDAK heritability models.

These analyses use Hybrid-Zero and Hybrid-GC, versions of SumHer that assign weights 1-p and p to the GCTA and LDAK heritability models, respectively. **a**, Average estimates of p from Hybrid-Zero for GCTA phenotypes (true p = 0), LDAK phenotypes (true p = 1) and hybrid phenotypes (true p = 0.5). **b**, Estimates of p from Hybrid-Zero for the 25 raw GWAS. Colors distinguish between the 13 WTCCC, 5 binary eMERGE and 7 quantitative eMERGE traits (black denotes the 25-trait average). A precise estimate of p was not possible for shingles (Segment 17), due to the trait having very low h^2_{SNP}. **c**, Estimates of p from Hybrid-GC for the 24 summary GWAS. Colors distinguish between the 9 binary and 15 quantitative traits (black denotes the 24-trait average). In all plots, vertical line segments mark 95% confidence intervals.

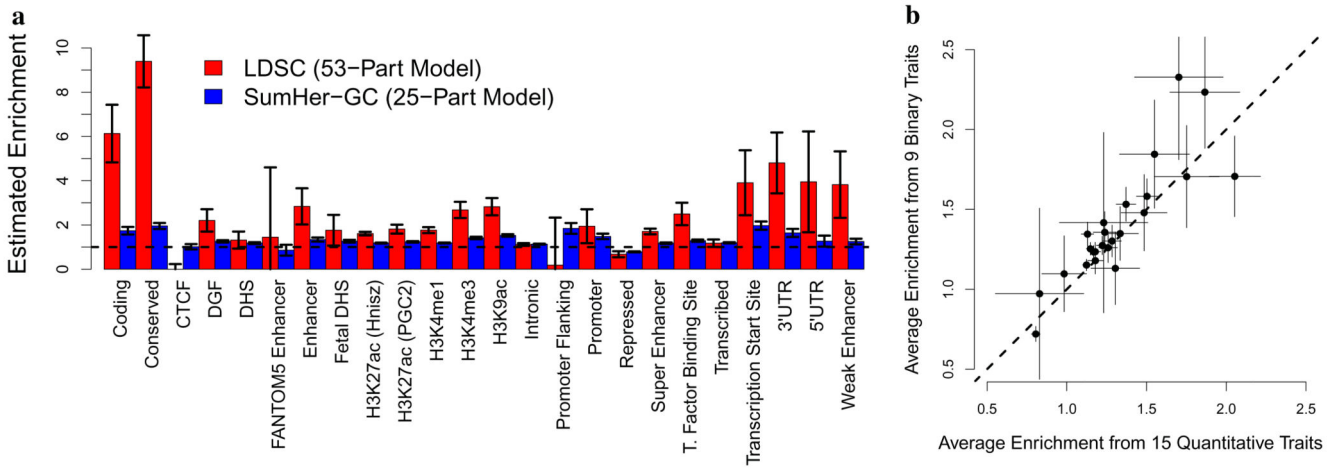

Figure 3. Functional enrichments across the 24 summary GWAS.

a, Average estimates of enrichments for the 24 functional categories from LDSC (red bars) and from SumHer-GC (blue bars). **b**, Average estimates of enrichments from SumHer-GC, based either on the 9 binary or on the 15 quantitative traits. In both plots, horizontal and vertical line segments mark 95% confidence intervals.

Europe PMC Funders Author Manuscripts

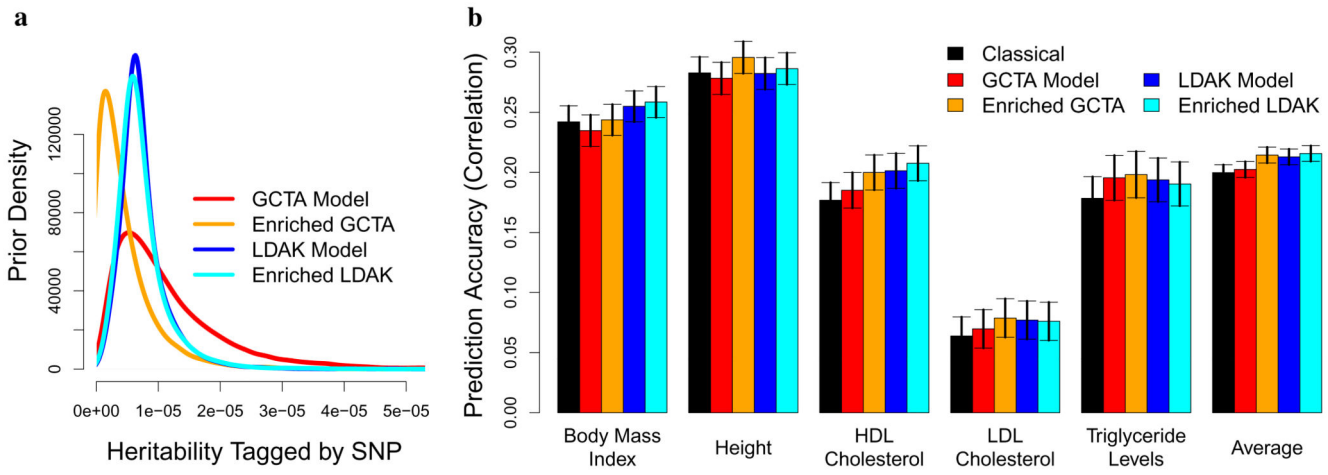

Figure 4. Prediction of five quantitative traits.

For each trait, we use data from the 24 summary GWAS to construct Bayesian polygenic risk scores (PRS) corresponding to four heritability models: GCTA, Enriched GCTA, LDAK and Enriched LDAK (see main text for details of each model). **a**, The distribution of $E[v^2_j]$, the expected heritability tagged by SNP j, corresponding to each heritability model. **b**, Prediction accuracy, measured as correlation between observed and predicted phenotypes in the (independent) eMERGE data, for the Classical PRS (effect sizes are frequentist estimates from single-SNP analysis) and for each of the four Bayesian PRS (effect sizes are posterior means). Vertical line segments mark 95% confidence intervals.

Europe PMC Funders Author Manuscripts

Table 1

Estimates of h^2_{SNP} and confounding bias for the 24 summary GWAS. Columns 2 & 3 report the average sample size and the genomic inflation factor (calculated using the published test statistics). Columns 4-11 report estimates of h^2_{SNP} and confounding from both LDSC and SumHer-GC (LDSC measures confounding via the intercept, 1 + A, while SumHer-GC uses the scaling factor, C). For binary traits, estimates of h^2_{SNP} have been converted to the liability scale, assuming the stated prevalence. Columns 12-15 report the number of significant loci based on the published test statistics, then after correction via genomic control, LDSC and SumHer-GC (dividing test statistics by the GIF, 1 + A and C, respectively).

			LDSC				SumHer-GC				No. of significant loci after dividing test statistics by			
Trait (disease prevalence, %)	**n**	**GIF**	**h^2_{SNP}**	**s.d.**	**1+A**	**s.d.**	**h^2_{SNP}**	**s.d.**	**C**	**s.d.**	**1**	**GIF**	**1+A**	**C**
Alzheimer's Disease46 (7.5)	54000	1.09	0.03	0.01	1.07	0.02	0.12	0.03	1.03	0.01	21	19	19	21
Coronary Artery47 (6)	79000	1.10	0.04	0.01	1.06	0.01	0.15	0.02	0.99	0.01	10	6	7	10
Crohn's Disease36 (0.5)	21000	1.14	0.15	0.03	1.08	0.01	0.47	0.06	0.97	0.02	64	52	58	64
Ever Smoked?48 (56)	74000	1.11	0.08	0.01	1.02	0.01	0.19	0.02	0.96	0.01	0	0	0	0
Inflammatory Bowel36 (0.7)	35000	1.17	0.09	0.02	1.13	0.01	0.33	0.03	0.98	0.01	78	59	65	80
Rheumatoid Arhritis49 (0.5)	58000	1.05	0.05	0.01	1.00	0.01	0.17	0.03	0.90	0.02	109	104	109	123
Schizophrenia50 (1)	82000	1.57	0.19	0.01	1.16	0.01	0.42	0.02	0.91	0.01	105	23	63	140
Type 2 Diabetes35 (8)	157000	1.17	0.08	0.01	1.07	0.01	0.23	0.02	0.95	0.01	38	25	32	42
Ulcerative Colitis36 (0.2)	27000	1.12	0.06	0.01	1.10	0.01	0.27	0.03	0.99	0.01	38	31	31	38
Bone Mineral Density37	33000	1.11	0.10	0.02	1.07	0.01	0.28	0.04	1.00	0.01	19	18	18	19
Body Mass Index25	230000	1.13	0.09	0.01	0.80	0.01	0.33	0.03	0.55	0.02	69	52	135	336
Depressive Symptoms38	161000	1.12	0.02	0.00	1.03	0.01	0.07	0.01	0.96	0.01	0	0	0	1
Fasting Glucose39	58000	1.08	0.05	0.01	1.04	0.01	0.14	0.03	0.99	0.01	22	20	20	23
Glycated Hemoglobin40	46000	1.04	0.02	0.01	1.03	0.01	0.10	0.02	0.99	0.01	10	10	10	10
HDL Cholesterol26	96000	1.03	0.07	0.03	1.04	0.07	0.50	0.09	0.68	0.03	130	122	121	216
Height41	246000	2.09	0.20	0.02	1.69	0.06	0.46	0.04	0.98	0.04	720	196	288	754
LDL Cholesterol26	91000	1.03	0.08	0.03	1.00	0.04	0.43	0.10	0.73	0.04	101	96	101	155
Menarche Age42	253000	1.66	0.15	0.01	1.21	0.02	0.32	0.02	0.89	0.02	289	111	190	354
Menopause Age43	69000	1.10	0.06	0.01	1.06	0.02	0.25	0.03	0.92	0.02	49	39	39	55
Neuroticism38	171000	1.26	0.06	0.01	1.06	0.01	0.17	0.02	0.90	0.02	10	4	7	18
Subjective Well-Being38	298000	1.16	0.02	0.00	1.03	0.01	0.04	0.00	0.97	0.02	0	0	0	0
Triglyceride26	92000	1.02	0.14	0.04	0.92	0.03	0.45	0.11	0.70	0.04	82	82	91	152
Waist-Hip Ratio44	142000	1.05	0.06	0.01	0.92	0.01	0.20	0.02	0.76	0.01	26	23	33	66
Years Education45	329000	1.54	0.07	0.00	1.11	0.01	0.20	0.01	0.83	0.01	70	13	46	148
Average	**121000**	**1.21**	**0.04**	**0.00**	**1.04**	**0.00**	**0.12**	**0.00**	**0.93**	**0.00**	**86**	**46**	**62**	**118**
Total											**2060**	**1105**	**1483**	**2825**